UNIVERSITY COLLEGE LONDON

DEPARTMENT OF STATISTICAL SCIENCE

FACULTY OF MATHEMATICAL AND PHYSICAL SCIENCES

A

T H E S I S

submitted for the Degree of

Doctor of Philosophy

of the University College London

by

Francesco Donat

Discrete Responses in Penalized Generalized Linear Models

UCL

DECEMBER MMXV

Declaration

I, Francesco Donat confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

.....

Francesco Donat

Bruxelles and London, December 18, 2015

Frankfurt am Main, April 18, 2016

Supervisors:	Dr Giampiero Marra
	Prof Thomas Fearn
Examiners:	Dr Gareth Peters (UCL)
	Dr Jochen Einbeck (Durham)
Thesis submission:	18/12/2015
Viva:	11/02/2016

Discrete Responses in Penalized Generalized Linear Models

Abstract: Generalized Linear Models (GLMs) are an important class of models that provide a unifying framework for the analysis and estimation of several types response variables, including discrete outcomes. This thesis discusses the representation, estimation and some inferential results of various models for categorical responses within a penalized GLM structure. In particular, a ridge-type penalty form is included to enforce certain properties of the functional form of the covariate-response relationship. Specifically, fully, non-parametric effects as well as smoothed spatial dependencies are shown to be all be represented through an appropriate combination of linear predictors and penalisation terms.

The emphasis of the thesis is on bivariate models for discrete responses that are commonly employed in cross-sectional studies to correct for the presence of direct unmeasured confounding and/or non-random sample selection issues. The former refers to a situation where both the response of interest and one of its relevant covariate are affected by a third variable, the confounder, which is either unobserved or not readily quantifiable by the researcher. The latter, instead, accounts for those instances where item non-response does not occur at random, but is driven by some underlying factors. In either case, not controlling for pertinent confounders may lead to detrimental effects in the estimates obtained, and standard estimators are usually inconsistent. Under certain conditions, bivariate models are proven to mend these issues.

The thesis shows how both types of models can be represented within a unified penalized GLM framework for discrete responses. Methodological advances are then provided towards two main research avenues: (i) the estimation of non-parametric covariate effects and smoothed spatial dependencies, and (ii) an improved flexibility achieved through the specification of copula functions for the idiosyncratic model components. In this way, several alternative dependence structures among the responses are also introduced. The extensive use of real datasets illustrates each situation in details and completes the analysis.

Keywords: Archimedean copulae; Bivariate system of equations; Ordinal responses; Penalized regression splines; Unobservables.

Acknowledgments

[05/12, London, UK] Waiting for a bus has been possibly the recurrent theme of my spare time here in the UK. So it seems natural to find myself in the situation of writing these few lines when I'm about to leave the country once again, waiting for a bus. Quoting an ancient and wise philosopher: "The best thing of traveling is to have the opportunity to sit down, smoke the pipe and write your last will". Ok, maybe these are not his exact words, nonetheless this idea can be easily translated for the PhD. I hereby solemnly affirm that the best thing of undertaking doctoral studies is to write the acknowledgments. And since an extensive reasoning would probably require more than what the next hundred pages would allow me, I will just mention an illustrative example. However absurd, a dissertation seems to me the best way for a depressed and unknown writer to get his literary works published, although only as part of a miserable acknowledgment section, and solely available from a university repository. But better in this way than leaving them written in a dusty copybook. Indeed any thesis title would sound much more appealing in the form:

> Discrete Responses in Penalized Generalized Linear Model with an Introductory Poem on Mr Robin Goodfellow dancing around the Pandora's Box.

Alex, you would have been so proud of that! But I'm not a writer, and I'm a terrible dancer (anyway you will never figure this out [wink]) with the poetic exuberance of William McGonagall. So what remains when the last journey begins and the white shores about to be reached? Perhaps not more than the desire to cross the Channel and the somehow childish melancholy of looking at the city lights through the windows of the overnight bus. *[Dear lady sitting in front of me, please, would you kindly put your seat in the upright position? And you, Dutches opposite the aisle, don't you realise how annoying you are at chatting from a three-row distance? Especially when I'm tipsy, have a headache and experiencing my epiphany...]*

[16/12, Strasbourg, FR] The Cathedral's clock is ticking and my time is running out: I'm afraid I have not the time to write everything I wanted. I better move to the customary thanking. The Svenska Akademien sincerely welcomes.

[16/12, Strasbourg, FR] My deepest gratitude goes certainly to my supervisor Dr Giampiero Marra, for his constant guidance and the many discussions we had during the past

three years. Much I've learned from him, but apparently the ability to write in plain English. Many other people have indirectly contributed, voluntary or not, to this thesis. In particular, with my fellow colleagues at the LSE I shared several ideas and opinions on life, politics, economics and, of course, statistics. I've discussed with them extensively about the possibility of whether to pursue a PhD or not: so thanks to Patrick, Marcel, Tianmiao "Marianne", Jay, Costanza and Lily for their advices, suggestions and the really good times we had together. I miss those moments. *[Apparently I'm chatting too loud, the clerk of the European Court of Human Rights has just intimated me silence during the hearing in the Grand Chamber.]*

[16/12, Somewhere along the way Strasbourg-Luxembourg-Bruxelles] The "On the road" (not-so-dream-)team gave me a strong support, more than just a drinking one (alas, not everyone drinks, despite they are all graduated!). Puppo the driver and Jimbo non-so-chi-o-cosa were simply great companions in many adventures around the world, while the then Hon Scabbia learned impressively how to deal with the disadvantages of having the "crazy one" as flatmate. The other one gave up throughout the process: not strong enough! Dr Mitch, since we moved to the UK *[wait, do you really think that traveling by train is better than the overnight bus? Have a ride on the Strasbourg-Bruxelles line first...]* you have been the main target of all my disappointments (many) and successes (relatively scarce). It was great to travel around the country to join you in parties, moving outs, hiking, visits, or just for the necessity of leaving the "great cesspool" for a while. Now wait in the US until I catch the flight from anywhere in continental Europe (Teutonic, perhaps?) to visit you once again.

The privileged red phone with the UCL Department of Economics (aka Skype) was similarly helpful and important to me. Thanks Cry for having been constantly available for doubts, complaints or simply for pointless chats on many matters of doubtful interest. Arguably, you have been so far "colleague" of mine for the longest time, since the very first year bachelor in Rome. From "esercicci" (do you remember her?), passing through the wonderful "cinesa" and the superlative "P...lloni bla bla bla". And how to forget Gloria, Carlotta, Mirko and Cats? Ah *bright college days, oh carefree days that fly.*

Complain, complain and again complain. About the supervisor, the students and the data that "don't work", shouting against the dissertation, wishing to shatter the computer and looking for a real job. Do the feelings sound familiar, Sara? But remind that there's the light eventually, and the end of the tunnel is close... Wow, I can't believe it's (almost) over!

I had a great time at the Department of Statistical Science of UCL, where I enjoyed many nice moments with good friends, *in primis* Rod, Anne-Marie, Beate and, more recently, Verena, Nayia and Pete. Rod and Bibi, thank you a lot for having hosted me in so many occasions throughout my teaching peregrinations from Brussels. I hope you will enjoy daily coffee breaks in your new EP-branded cup(s).... I'm also indebted to the Economic Governance Support Unit of the European Parliament, in particular my managers Marcel Magnus and Kajus Hagelstam, for having granted me the necessary time to complete the thesis and go to London regularly during my scholarship. Thanks Alice for your support and the many laughs on statistics and economists playing statisticians in European Institutions.

As always, people in the shadow share a huge part in any achievement. Their support goes way beyond any particular action they might have taken, or words they might have spoken: *to thee we sing with our glasses raised on high.*

A final word goes to some of the mentors and teachers of economics, econometrics and statistics I had at Tor Vergata, LSE and UCL: among them Gustavo Piga, Paolo Paesani, Tommaso Proietti and Qiwei Yao. If anything good academically ever came out of me is largely because of their passion, dedication and motivation during my younger years. I hope to have transmitted at least a little share of those every day to my students. Richard Blundell, Andrew Chesher and Adam Rosen had instead the merits of having introduced me to fascinating areas of econometric research I wish I can work on some day. Perhaps my next life, eh?

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Prelude

This thesis concerns the development, representation and estimation of flexible discrete response regression models which account for the role that, at different levels, unobserved variables may have on an outcome of primary interest for the researcher. This is achieved by specifying appropriately a bivariate system of non-linear equations. Several models are discussed, and their functioning demonstrated throughout relevant empirical applications. The proposed models are finally described as special instances of a unifying regression framework.

Under a methodological point of view, my representation extends the class of Generalized Additive Models (GAMs; Hastie and Tibshirani, 1986, 1990) to the multivariate dimension in the spirit of Vector GAMs (VGAMs), originally introduced by Yee and Wild (1996). The proceeding work, however, is distinguished from the latter study for the inclusion of a computationally stable and efficient way to perform smoothing parameter estimation, as based on the results of Wood (2004). Moreover, a novel semi-parametric triangular structure is considered, whose statistical properties have not been studied in the literature yet. The main features of the proposed models include an enhanced representation of the functional form of the covariate effects through penalized regression splines and smoothed spatial effects, as well as the analysis of various dependence structures by means of Archimedean copulae. All the necessary computational routines are also made available through the functions `SemiParCLM` and `CopulaCLM` for the `R` environment.

This introduction is structured as follows: I next introduce univariate semi-parametric regression models for discrete outcomes, and subsequently review the representation of unknown smooth curves through penalized regression splines. I finally describe the unmeasured confounding problem in the simplified context of continuous responses to motivate the class of models discussed in the thesis.

1.1.1 Discrete Responses within a GLM Representation

Traditionally, discrete data are dealt as instances of the class of Generalized Linear Models (GLMs, Nelder and Wedderburn, 1972), a generic representation that connects the expected value of a response of interest, $Y \in \mathcal{Y}$, conditioned to a set of explanatory variables, $\boldsymbol{x}$, through a known *link function*, $g : \mathcal{Y} \longrightarrow \mathbb{R}$. In the univariate case, this corresponds to specifying a model of the form

$$\mathbb{E}[Y|\boldsymbol{X} = \boldsymbol{x}] = g^{-1}(\eta), \tag{1.1}$$

where $\eta := \mathbf{x}^\top \boldsymbol{\beta} \in \mathbb{R}$ is usually denominated the *linear predictor*. Different models involve distinct response types and link functions. For instance, a binary outcome would set $\mathcal{Y} = \{0, 1\}$, for which a cumulative distribution function is usually employed as link. This gives rise to well-known models like the probit, logit, or the complementary log-log, wherever a Standard Normal, a Logistic or a log-Weibull are used for g^{-1}, respectively.

Consider now n realisations of the response variable, $\mathbf{y} = (\mathrm{y}_1, \ldots, \mathrm{y}_n)^\top$, and let $\mathbf{Z}$ be the corresponding model design matrix, then any GLM is fully determined by the pair $(g, \mathbf{Z})$. In particular, linear models for continuous responses, as well as for categorical and count outcomes can all be obtained as specific instances of (1.1). Although this class of models is remarkably rich in terms of the distributions allowed, and a rather elegant theory of estimation and inference exists, the pre-determined (linear) parametric form of the covariate effects may in practice limit its relevance in applied research.

A higher degree of flexibility – but still relatively computationally cheap – can be achieved upon extending η to account for a non-parametric specification of possible non-linear effects of the continuous covariates v_l's, $l = 1, \ldots, L$. Namely, for some smooth functions $s_l : \mathbb{R} \longrightarrow \mathbb{R}$ to be estimated, one can consider instead a model consisting of

$$\eta = \mathbf{x}^\top \boldsymbol{\beta} + s_1(\mathrm{v}_1) + \cdots + s_l(\mathrm{v}_l) + \cdots + s_L(\mathrm{v}_L). \tag{1.2}$$

The above representation defines the so-called GAM, which is effectively a GLM in which the linear effects of the continuous covariates are replaced by the sum of some unknown smooth functions. In the thesis, these are estimated using penalized regression splines. Therefore, a GAM allows for a data-driven representation of the covariate-response relationship, and does not require the imposition of any pre-determined functional form of the covariate effects on the outcome of interest. In the proceeding discussion, any form like (1.2), which include

both fully and non-parametric covariate effects, will be referred to as *semi-parametric*.

Since the main focus of the thesis is the analysis of discrete data, it is necessary to incorporate them within a GLM framework in a comprehensive way. To this end, after giving some preliminary definitions used throughout the work, I follow and extend to the present context the approach outlined by Peyhardi et al. (2014) for the univariate case.

Let Y be a random variable whose support is the discrete set $\mathcal{Y} \equiv \mathcal{K} = \{1, \dots, K\} \subset \mathbb{N}$, with $K < \infty$. For $\#(\mathcal{K}) = 2$, Y is called *dichotomous* (or binary) whereas, in the case of $\#(\mathcal{K}) > 2$, the response variable is said *polychotomus*. The elements $k \in \mathcal{K}$ are referred to as the *levels* (or categories) of Y, and may represent both quantitative and qualitative measurements. Following the classification introduced by Stevens (1946), it is also meaningful to distinguish between polychotomous variables as measured on the *nominal* and *ordinal scale*. The former differentiates items based only on the categories they belong to, whereas the latter allows also for a rank order by which the realisations of Y can be sorted. In this case, the relative degree of difference between them cannot be interpreted. This is formalised by assuming that the support of Y, $(\mathcal{K}, \preceq)$, where $\preceq$ is a binary relation, is totally ordered and the response is said to be *ordinal*. Specifically:

Definition 1. *The set $\mathcal{K}$ is* totally ordered *under $\preceq$ if, for all $k, \bar{k}, \tilde{k} \in \mathcal{K}$, the following statements hold:*

1. $k \preceq k$ (reflexivity),
2. *if* $k \preceq \bar{k}$ *and* $\bar{k} \preceq k$, *then* $k = \bar{k}$ (antisymmetry),
3. *if* $k \preceq \bar{k}$ *and* $\bar{k} \preceq \tilde{k}$, *then* $k \preceq \tilde{k}$ (transitivity),
4. $k \preceq \bar{k}$ *or* $\bar{k} \preceq k$ (totality).

The above definition implies that, for any value the random variable Y can take, there exists a unique way in which the elements in its support can be related under $\preceq$, and each of them is comparable with respect to all the others. The exact interpretation of "$\preceq$" would clearly depend on the variable Y considered. For example, for Y being "severity injury in one-vehicle accident", the binary relation can be translated to as "less severe than".

Most models for discrete data can be motivated by the existence of a continuous latent variable Y^*, and the specification of suitable assumptions connecting it to the levels of Y. With reference to ordinal polychotomous variables, the equivalence linking the latent

response to the ordinal manifest one is given by

$$\{Y = k\} \Longleftrightarrow \{c_{k-1} < Y^* \leq c_k\},$$

where $c_1 \leq c_2 \leq \cdots \leq c_K = \infty$, $c_0 := c_{1-1} = -\infty$, are term cut points or threshold parameters. Notice that, wherever the c_k's are the only parameters in the linear predictor depending on the ordered levels of Y, namely

$$\eta_k := c_k - \mathbf{x}^\top \boldsymbol{\beta} - s_1(\mathrm{v}_1) - \cdots - s_L(\mathrm{v}_L), \tag{1.3}$$

it holds that $\{Y \leq k\} \Longleftrightarrow \{Y^* \leq \eta_k\}$. Hence, upon defining $\pi_k := \mathbb{P}[Y = k|\boldsymbol{X} = \boldsymbol{x}]$, it is possible to re-write model (1.1) for ordinal polychotomous responses as

$$r(\pi_k) = F_1(\eta_k), \tag{1.4}$$

where $F_1 \equiv g$ is any 1-variate cdf, and $r : [0,1] \longrightarrow [0,1]$ a map characterising the type of the outcome. For ordinal polychotomous responses, for instance, $r(\pi_k) := \pi_1 + \cdots + \pi_k$, and the corresponding specification of (1.4) has been termed *cumulative link model* (CLM) by McCullagh (1980). Notice that in the proceeding of the thesis I will keep using the simplified notation of $r(\pi_k)$ *in lieu* of the more precise $r(\pi_1, \ldots, \pi_k)$ to emphasise the common structure of GLM models for discrete responses.

Example 1. According to model (1.4) above, one can represent a logit regression by imposing the map $r_k : \pi_k \mapsto \pi_k$, so that

$$\pi_k = \mathrm{logistic}(\eta_k), \qquad k \in \{0, 1\}.$$

In a similar fashion, the ordered probit model in the CLM form is simply

$$\pi_1 + \cdots + \pi_k = \Phi(\eta_k).$$

□

Recently, Peyhardi et al. (2014) studied a general form of (1.4) for a fully parametric class of models, and defined uniquely any GLM for categorical variables in terms of the triplet $(r, F_1, \mathbf{Z})$.

Some comments are in order. The right-hand side of (1.4) is *model specific* as it depends

only on the functional form of the covariates on the response and the link function employed. The left-hand side, instead, is peculiar to the support of Y (and so to its type). To appreciate this difference, observe that a semi-parametric model – as expressed by (1.4) – and its parametric counterpart differ only from the specification of η_k. Therefore, methodological advances accounting for this modularity would in principle allow for the creation of a generic framework dealing with several model specifications of (1.4). This idea has been constantly used throughout the thesis, and explicitly targeted in some generality in the first part of Chapter 4.

1.2 Penalized Regression Splines Approximation

Smooth functions in (1.3) need to be appropriately represented and estimated. This is achieved here by penalized regression splines, a method that comprises a wide spectrum of smoothers deriving from the choice of different basis functions, type of penalties, amount and location of knots. In what follows, a number of techniques are reviewed and used to illustrate a univariate discrete response model with linear predictor specified as in (1.3).

Complex covariate-response relationships beyond a purely polynomial functional form – as encoded in the maps s_l's – can be represented by defining a set of H_l+1 knot points, v^*_{l,h_l}, and completing a polynomial model with a *truncated polynomial basis function*. Namely

$$s_l(\mathrm{v}_l) = \bar{\delta}_{l,0} + \bar{\delta}_{l,1}\mathrm{v}_l + \cdots + \bar{\delta}_{l,q}\mathrm{v}_l^q + \sum_{h_l=1}^{H_l+1} \delta_{l,h_l}(\mathrm{v}_l - \mathrm{v}^*_{l,h_l})_+^q = \boldsymbol{\delta}_l^\top \boldsymbol{b}_l, \qquad l = 1, \ldots, L,$$

where $(\mathrm{v}_l - \mathrm{v}^*_{l,h_l})_+ := \max\{0, \mathrm{v}_l - \mathrm{v}^*_{l,h_l}\}$, $\boldsymbol{\delta}_l := (\bar{\delta}_{l,0}, \ldots, \bar{\delta}_{l,q}, \delta_{l,1}, \ldots, \delta_{l,H_l+1})^\top$ and $\boldsymbol{b}_l := (1, \mathrm{v}_l, \ldots, \mathrm{v}_l^q, (\mathrm{v}_l - \mathrm{v}^*_{l,1})_+^q, \ldots, (\mathrm{v}_l - \mathrm{v}^*_{l,H_l+1})_+^q)^\top$. Assume now that n observations are available and let M denote the number of regressors, then it is possible to define the $n \times M$ matrix $\mathbf{X} := (\mathbf{x}_1 | \cdots | \mathbf{x}_n)^\top$ and, accordingly, $\mathbf{B}_l := (\boldsymbol{b}_{l,1} | \cdots | \boldsymbol{b}_{l,n})^\top$, so that $\mathbf{B} := (\mathbf{B}_1, \ldots, \mathbf{B}_l, \ldots, \mathbf{B}_L)$ and $\boldsymbol{\delta} := \mathrm{vec}(\boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_l, \ldots, \boldsymbol{\delta}_L)$. Hence, for $\boldsymbol{\eta} := (\eta_i)_{i=1}^n \in \mathbb{R}^n$ and $\boldsymbol{c} := (c_i)_i \in \mathbb{R}^n$, linear predictors in (1.3) can be expressed equivalently by

$$\boldsymbol{\eta} = \boldsymbol{c} - \mathbf{X}\boldsymbol{\beta} - \mathbf{B}\boldsymbol{\delta} = \mathbf{Z}\boldsymbol{\beta},$$

where $\mathbf{Z} := (\mathbb{I}, -\mathbf{X}, -\mathbf{B})$, $\mathbb{I} := (\mathbb{1}_{\mathrm{y}_i = k})_{i,k} \in \{0,1\}^{n \times K-1}$ and $\boldsymbol{\vartheta} := \mathrm{vec}(\boldsymbol{c}_k, \boldsymbol{\beta}, \boldsymbol{\delta})$ with $\boldsymbol{c}_k := (c_k)_k \in \mathbb{R}^{K-1}$. The above model is now in a purely parametric form, and hence estimable as such. For discrete responses, it can be proved that the score and Hessian matrix of the

corresponding log-likelihood function, $\ell(\boldsymbol{\vartheta}|\mathbf{y}, \mathbf{Z})$, can be written as

$$\nabla_{\boldsymbol{\vartheta}}\ell(\boldsymbol{\vartheta}|\cdot) = \mathbf{D}^\top\mathbf{u} \quad \text{and} \quad \nabla_{\boldsymbol{\vartheta}\boldsymbol{\vartheta}^\top}\ell(\boldsymbol{\vartheta}|\cdot) = \mathbf{D}^\top\mathbf{W}\mathbf{D},$$

with $\mathbf{D}$, $\mathbf{u}$ and $\mathbf{W}$ arrays appropriately defined (please refer to Section 4.4 for the most general form they can take for the purpose of the thesis). Remarkably, each iteration of the Newton-Raphson (or Fisher Scoring) algorithm to log-likelihood maximisation solves the generalized least squares problem

$$\boldsymbol{\vartheta}^{[\alpha+1]} = \arg\min_{\boldsymbol{t}} \left\|\mathbf{W}^{1/2}(\mathbf{z} - \mathbf{D}\boldsymbol{t})\right\|^2\Big|_{\boldsymbol{\vartheta}=\boldsymbol{\vartheta}^{[\alpha]}} = (\mathbf{D}^\top\mathbf{W}\mathbf{D})^{-1}\mathbf{D}^\top\mathbf{W}\mathbf{z}\Big|_{\boldsymbol{\vartheta}=\boldsymbol{\vartheta}^{[\alpha]}},$$

where $\mathbf{z} := \mathbf{D}\boldsymbol{\vartheta} - \mathbf{W}^{-1}\mathbf{u}$ is termed *pseudo-data vector*. These quantities have been left generic intentionally as their exact definition will crucially depend on the map r, as well as the link function used. This will appear clear from the general derivations of Section 4.4.

The estimation approach just outlined is commonly referred to as *regression spline smoothing*. Although appealing, this method is effectively limited by the issue of selecting the number and location of the knots, for which proposed methodologies are usually rather complicated and computationally intensive (e.g., Friedman and Silverman, 1989 or Stone et al., 1997). A valuable alternative to overcome these drawbacks is to impose a penalty on the spline coefficients. To this end, a researcher can choose a sufficiently large number of knots points and control for over-fitting by constraining the coefficients as

$$\min_{\boldsymbol{t}} \left\|\mathbf{W}^{1/2}(\mathbf{z} - \mathbf{D}\boldsymbol{t})\right\|^2 \quad \text{subject to} \quad \|\boldsymbol{\delta}\|^2 \leq c$$

for some constant $c \geq 0$. Using a Lagrange multiplier argument, it can be shown that the above optimisation problem can be written as

$$\min_{\boldsymbol{t}} \left\{\left\|\mathbf{W}^{1/2}(\mathbf{z} - \mathbf{D}\boldsymbol{t})\right\|^2 + \boldsymbol{\delta}_{\boldsymbol{\lambda}}^\top\boldsymbol{\delta}_{\boldsymbol{\lambda}}\right\} = \min_{\boldsymbol{t}} \left\{\left\|\mathbf{W}^{1/2}(\mathbf{z} - \mathbf{D}\boldsymbol{t})\right\|^2 + \boldsymbol{t}^\top\mathbf{S}_{\boldsymbol{\lambda}}\boldsymbol{t}\right\},$$

where $\mathbf{S}_{\boldsymbol{\lambda}} := \text{diag}(\mathbf{0}, (\mathbf{S}_l)_l)$, $\mathbf{S}_l := \lambda_l\mathbf{I}_{H_l}$ and λ_l is a non-negative real number for any l. Hence it holds that

$$\boldsymbol{\vartheta}^{[\alpha+1]} = (\mathbf{D}^\top\mathbf{W}\mathbf{D} + \mathbf{S}_{\boldsymbol{\lambda}})^{-1}\mathbf{D}^\top\mathbf{W}\mathbf{z},$$

which is a ridge-type estimator with penalisation depending on a *smoothing parameter* λ_l. In particular, the larger its magnitude, the more the estimates shrink towards a polynomial fit,

whereas $\lambda_l \to 0$ results in wiggly values. This *penalized regression splines* idea traces back at least to the works of Wahba (1980) and Parker and Rice (1985). Although the derivations above follow from the employment of a L_2 penalty, different degrees of $\|\boldsymbol{\delta}\|^p$, $p = 1, 2, \ldots$, are also plausible. They are not discussed in this work because some distributional results and subsequent algorithms strongly rely on the imposition of a ridge-type penalty (e.g. Section 2.3.4).

Despite its simple construction and use, the truncated polynomial basis just described suffers from a number of numerical instabilities. In fact, for $\boldsymbol{\lambda} := (\lambda_l)_l \in \mathbb{R}_+^L$ close to $\mathbf{0}$, the inversion of matrix $(\mathbf{D}^\top \mathbf{W} \mathbf{D} + \mathbf{S}_{\boldsymbol{\lambda}})$ may rise computational concerns. Acknowledging this fact, I next review some competing and stable alternatives in the context of penalized regression splines. They are all implemented and usable for all the models introduced in this thesis.

1.2.1 Penalized B-splines

Eilers and Marx (1996) proposed P-splines as a penalized version of the the B-splines basis described by de Boor (1978). These basis functions are strictly local since each of them is non-zero only between $q + 3$ adjacent knots, where $q + 1$ denotes the order of the basis. Under this convention, a *cubic spline* corresponds to the setting of $q = 2$.

The definition of a H_l-parameter B-spline basis follows from the location of $H_l + q + 1$ (usually equally spaced) knot points $\mathrm{v}_{l,1}^* < \mathrm{v}_{l,2}^* < \cdots < \mathrm{v}_{l,H_l+q+1}^*$, so that the spline function is evaluated over the closed interval $[\mathrm{v}_{l,q+2}^*, \mathrm{v}_{l,H_l}^*]$. The $(q+1)$-th order spline is then described by

$$s_l(\mathrm{v}_l) = \sum_{h_l=1}^{H_l} B_{l,h_l}^q(\mathrm{v}_l)\delta_{l,h_l},$$

for $\delta_{l,h_l} \in \boldsymbol{\delta}_l$, and the bases recovered recursively as

$$B_{l,h_l}^q(\mathrm{v}_l) = \frac{\mathrm{v}_l - \mathrm{v}_{l,h_l}^*}{\mathrm{v}_{l,h_l+q+1}^* - \mathrm{v}_{l,h_l}^*} B_{l,h_l}^{q-1}(\mathrm{v}_l) + \frac{\mathrm{v}_{l,h_l+q+2}^* - \mathrm{v}_l}{\mathrm{v}_{l,h_l+q+1}^* - \mathrm{v}_{l,h_l+1}^*} B_{l,h_l+1}^{q-1}(\mathrm{v}_l),$$

for $h_j = 1, \ldots, H_l$ and $B_{l,h_l}^{-1} = \mathbb{1}_{\mathrm{v}_{l,h_l}^* \le \mathrm{v}_l < \mathrm{v}_{l,h_l+1}^*}$. P-splines are usually completed with a difference penalty applied to the parameters δ_{l,h_l}'s to control for their wiggliness. For example, if the researcher wishes to penalise the squared difference between adjacent δ_{l,h_l} values, the

penalty would take the form

$$\mathcal{P}_{\lambda_l} = \lambda_l \sum_{h_l=1}^{H_l-1} (\delta_{l,h_l+1} - \delta_{l,h_l})^2 = \delta_{l,1}^2 - 2\delta_{l,1}\delta_{l,2} + 2\delta_{l,2}^2 - 2\delta_{l,2}\delta_{l,3} + \cdots + \delta_{l,H_l}^2 \in \mathbb{R},$$

or, equivalently,

$$\mathcal{P}_{\lambda_l} = \lambda_l \boldsymbol{\delta}_l^\top \begin{bmatrix} 1 & -1 & & & \\ -1 & 2 & -1 & & \\ & -1 & 2 & -1 & \\ & & \ddots & \ddots & \ddots \end{bmatrix} \boldsymbol{\delta}_l.$$

1.2.2 Thin-plate Regression Splines

P-splines are useful in practice, but they suffer from some criticisms. Specifically, they still require to choose knots' location (a fact that introduces a further degree of subjectivity), and it is not clear to what extent they are better or worse than any other basis that might be used. Thin-plate spline (Duchon, 1977) and their low-rank approximation (Wood, 2003) offer a solution to both these issues.

Assume that n observations are available $(\bar{\mathrm{y}}_{l,i}, \mathbf{v}_{l,i})_i$, where each $\mathbf{v}_{l,i}$ is a d-dimensional vector, $d \leq n$, and

$$s_l(\mathbf{v}_{l,i}) := \bar{\mathrm{y}}_{l,i} - \varepsilon_{l,i}.$$

The idea of thin-plate spline smoothing is to estimate s_l by

$$\widehat{s}_l := \arg\min_{\boldsymbol{h}_l} \|\bar{\mathbf{y}}_l - \boldsymbol{h}_l\|^2 + \lambda_l J_{qd}(h_l), \tag{1.5}$$

where $\bar{\mathbf{y}}_l := (\bar{\mathrm{y}}_{l,i})_i$ collects the available data, whereas $\boldsymbol{h}_l := (h_l(\mathbf{v}_{l,i}))_i$ is a vector of unknowns, and J_{qd} a penalty functional measuring the wiggliness of the map h_l defined as

$$J_{qd} := \int \cdots \int_{\mathbb{R}^d} \sum_{\nu_1+\cdots+\nu_d=q} \frac{q!}{\nu_1! \cdots \nu_d!} \left(\frac{\partial^q h_l}{\partial v_1^{\nu_1} \cdots \partial v_d^{\nu_d}} \right)^2 \mathrm{d}v_1 \ldots \mathrm{d}v_d. \tag{1.6}$$

For instance, a smooth of only one predictor and with wiggliness measured using second derivatives would set $d = 1$ and $q = 2$ in the above. In this case, (1.6) reduces to

$$J_{21} = \int_{\mathbb{R}} \left(\frac{\partial^2 h_l}{\partial v_1^2} \right)^2 \mathrm{d}v_1,$$

which is the kind of penalisation term used constantly throughout the thesis. Under the

technical assumption that $2q > d$, the solution to problem (1.5) has the closed form

$$\widehat{s}_l = \sum_{i=1}^{n} \delta_{l,i} b_{qd}(\|\mathbf{v}_l - \mathbf{v}_{l,i}\|) + \sum_{j=1}^{Q} \alpha_{l,j} \phi_{l,j}(\mathbf{v}), \quad Q = \begin{pmatrix} q+d-1 \\ d \end{pmatrix}, \tag{1.7}$$

where $\boldsymbol{\delta}_l := (\delta_{l,i})_i$ and $\boldsymbol{\alpha}_l := (\alpha_{l,i})_i$ are unknown parameter vectors to be estimated. The former is subject to the constraint $\mathbf{T}_l^\top \boldsymbol{\delta}_l = \mathbf{0}$, with $\mathbf{T}_l := (\phi_{l,j}(\mathbf{v}_{l,i}))_{i,j}$. Notice that the functions $\phi_{l,i}$'s span the space of polynomials for which $J_{qd} = 0$, the so-called *null space* of J_{qd}: those functions are then considered to be completely smooth. The exact form of the remaining bases b_{qd}'s can be found in Wood (2006). Upon defining the matrix $\mathbf{E}_l := b_{qd}(\|\mathbf{v}_{l,i} - \mathbf{v}_{l,j}\|)_{i,j}$ the spline fitting problem becomes

$$\min_{\boldsymbol{\delta}_l, \boldsymbol{\alpha}_l} \|\mathbf{y}_l - \mathbf{E}_l \boldsymbol{\delta}_l - \mathbf{T}_l \boldsymbol{\alpha}_l\|^2 + \lambda \boldsymbol{\delta}_l^\top \mathbf{E}_l \boldsymbol{\delta}_l \quad \text{subject to} \quad \mathbf{T}_l^\top \boldsymbol{\delta}_l = \mathbf{0}. \tag{1.8}$$

It is worth stressing that the thin-plate spline $\widehat{s}_l$ has a number of attractive properties that make it close to an ideal smoother. In fact, (1.5) defines exactly what is meant by smoothness, and how much weight is needed to balance data structure and smoothness. A clear disadvantage, however, is the high computational cost of $\mathcal{O}(n^3)$ operations required to form $\widehat{s}_l$ by (1.7). To overcome this limitation, Wood (2003) proposed a low-rank approximation to thin-plate bases.

The intuitive idea of this approach is to truncate the space of the wiggly components of the thin-plate spline, namely those corresponding to the regression parameter $\boldsymbol{\delta}_l$, while leaving the $\boldsymbol{\alpha}_l$ components unchanged. This is achieved by taking the spectral decomposition of $\mathbf{E}_l = \mathbf{Q}_l \boldsymbol{\Lambda}_l \mathbf{Q}_l$, with $\boldsymbol{\Lambda}_l$ re-arranged such that each eigenvalue $|\Lambda_{l,i,i}| \geq |\Lambda_{l,i+1,i+1}|$, $i = 1, \ldots, n-1$, and the columns of $\mathbf{Q}_l$ are the corresponding eigenvectors. Let then $\mathbf{Q}_{l,k}$ be the matrix consisting of the first k columns of $\mathbf{Q}_l$, and $\boldsymbol{\Lambda}_{l,k}$ denote the top left $k \times k$ submatrix of $\boldsymbol{\Lambda}_l$. Hence, by restricting $\boldsymbol{\delta}_l$ to the column space of $\mathbf{Q}_{l,k}$, namely by setting $\boldsymbol{\delta}_l = \mathbf{Q}_{l,k} \boldsymbol{\delta}_{l,k}$ (in which case $\boldsymbol{\delta}_{l,k} = \mathbf{Q}_{l,k}^\top \boldsymbol{\delta}_l$), problem (1.8) is re-stated as

$$\min_{\boldsymbol{\delta}_{l,k}, \boldsymbol{\alpha}_l} \|\mathbf{y}_l - \mathbf{Q}_{l,k} \boldsymbol{\Lambda}_{l,k} \boldsymbol{\delta}_{l,k} - \mathbf{T}_l \boldsymbol{\alpha}_l\|^2 + \lambda \boldsymbol{\delta}_{l,k}^\top \boldsymbol{\Lambda}_{l,k} \boldsymbol{\delta}_{l,k} \quad \text{subject to} \quad \mathbf{T}_l^\top \mathbf{Q}_{l,k} \boldsymbol{\delta}_{l,k} = \mathbf{0}.$$

Moreover, by taking the QR decomposition of $\mathbf{Q}_{l,k}^\top \mathbf{T}_l$, it is possible to find an orthogonal column basis $\overline{\mathbf{Z}}_{l,k}$ such that $\mathbf{T}_l^\top \mathbf{Q}_{l,k} \overline{\mathbf{Z}}_{l,k} = \mathbf{0}$. Therefore, upon restricting $\boldsymbol{\delta}_{l,k}$ to this space, $\boldsymbol{\delta}_{l,k} = \overline{\mathbf{Z}}_{l,k} \widetilde{\boldsymbol{\delta}}_l$, the corresponding unconstrained rank k approximation to the smoothing spline

becomes

$$\min_{\widetilde{\boldsymbol{\delta}}_l,\boldsymbol{\alpha}_l} \|\mathbf{y}_l - \mathbf{Q}_{l,k}\boldsymbol{\Lambda}_{l,k}\overline{\mathbf{Z}}_{l,k}\widetilde{\boldsymbol{\delta}}_l - \mathbf{T}_l\boldsymbol{\alpha}_l\|^2 + \lambda\widetilde{\boldsymbol{\delta}}_l^\top\overline{\mathbf{Z}}_{l,k}^\top\boldsymbol{\Lambda}_{l,k}\overline{\mathbf{Z}}_{l,k}\widetilde{\boldsymbol{\delta}}_l.$$

This has computational cost of $\mathcal{O}(k^3)$. Finally, once the model is fitted, the thin-plate spline can be evaluated by plugging $\boldsymbol{\delta}_l = \mathbf{Q}_{l,k}\boldsymbol{\delta}_{l,k} = \mathbf{Q}_{l,k}\overline{\mathbf{Z}}_{l,k}\widetilde{\boldsymbol{\delta}}_l$ in (1.7). The use of $\mathbf{Q}_{l,k}$ is pivotal for the approximation used, and is optimal in the sense that it minimises simultaneously both the changes in the fitted values of the spline, given by $\mathbf{E}_l\widehat{\boldsymbol{\delta}}_l + \mathbf{T}_l\boldsymbol{\alpha}_l$, and the shape of the estimated curve, $\widehat{\boldsymbol{\delta}}_l^\top\mathbf{E}_l\widehat{\boldsymbol{\delta}}_l$. Mathematically, the worst possible changes after the truncation are given by

$$\widehat{e}_{l,k} = \max_{\boldsymbol{\delta}_l\neq\mathbf{0}} \frac{\|(\mathbf{E}_l - \widehat{\mathbf{E}}_{l,k})\boldsymbol{\delta}_l\|}{\|\boldsymbol{\delta}_l\|} \quad \text{and} \quad \widetilde{e}_{l,k} = \max_{\boldsymbol{\delta}\neq\mathbf{0}} \frac{\boldsymbol{\delta}_l^\top(\mathbf{E}_l - \widetilde{\mathbf{E}}_{l,k})\boldsymbol{\delta}_l}{\|\boldsymbol{\delta}_l\|^2},$$

where $\widehat{\mathbf{E}}_{l,k} := \mathbf{E}_l\mathbf{Q}_{l,k}\mathbf{Q}_{l,k}^\top$ and $\widetilde{\mathbf{E}}_{l,k} := \mathbf{Q}_{l,k}^\top\mathbf{Q}_{l,k}\mathbf{E}_l\mathbf{Q}_{l,k}\mathbf{Q}_{l,k}^\top$. Wood (2003) proved that the minima of $\widehat{e}_{l,k}$ and $\widetilde{e}_{l,k}$ are indeed attained whenever $\mathbf{Q}_{l,k}$ is employed; we refer the interested reader to his paper for a detailed reasoning of this result.

Because of their desirable properties, penalized thin-plate regression splines are set as the default options in the `R` computational routines `SemiParCLM` and `CopulaCLM` accompanying this thesis. Nonetheless, P- and cubic regression splines are also estimable by specifying `"ps"` and `"cr"`, respectively, in the definition of the smooth model components.

1.3 Modelling Unobservables Through Systems of Equations

The methods introduced in this thesis are motivated by the modelling of several ways in which unobserved or unmeasured explanatory variables may affect a given outcome of interest. The next chapters will detail specific instances in which this may occur in applications. Specifically, they will deal with (a) direct unmeasured confounding, (b) the omission of common factors in the joint analysis of response variables, and (c) non-random-sample selection of individuals into (or out of) the relevant sample. Empirical illustrations are provided for each cases (a)-(c). I briefly motivate here the first instance under simplified assumptions, whereas I refer directly to Chapters 3 and 4, respectively, for an account of the others.

1.3.1 Direct Unmeasured Confounding

In regression analysis some assumptions are needed to assure that the estimates obtained have the statistical properties of being unbiased and consistent. However, applied research

often involves situations where these conditions are not met. For example, in traditional linear models, one fundamental assumption requires the regressors to be asymptotically uncorrelated with the stochastic model components. When this does not occur, the so-called endogeneity emerges (e.g., Wooldridge, 2002, Ch. 5).

I consider here the regression of a response of primary interest, Y_2, on some measured covariates, also called the treatments, where a background variable is explanatory to Y_2 and to one of its directly explanatory variables, Y_1. This situation is referred to as *direct confounding effect*. Problems arise when the researcher fails to adjust for pertinent confounders as they might be either unknown or not readily quantifiable. When this is the case, the confounding effect induces endogeneity and the use of standard estimators typically yields inconsistent estimates. Hence, a major concern when estimating treatment effects is how to account for unmeasured confounders. A common approach to deal with them are instrumental variables (IVs). This method postulates the existence of some observable variables not included in the regression equation that are uncorrelated with the error term but partially correlated with the confounded variables. When the above conditions are satisfied, these variables are termed instrument candidates for the confounded variables. IV estimation can yield consistent parameter estimates and can be used in any kind of analysis in which unmeasured confounding is suspected to be present.

In this framework, unobservables are meant to be relevant factors which are omitted from the analysis for any reason. In general, as Angris and Krueger (2001) pointed out, if they could be measured and held constant in a regression, the bias from omitted variables could be eliminated. In experimental studies one solution to the problem is the assignment of the relevant treatment to the subjects via a randomisation mechanism, whose functioning is independent of any other factor (on this point also Frosini, 2006). When this is not feasible – for ethical or legal reasons, say, or in observational studies in general – consistent estimates can still be obtained by exploiting a certain degree of exogenous variation in the potentially endogenous treatments by means of instrumental variables. In fact, they solve the unmeasured confounding problem by splitting the variability in the endogenous predictors in two parts, one of which is uncorrelated with the confounder. The method then uses only this part to obtain consistent estimates of the effects of the variables of interest.

The following section aims at giving an account of the concepts outlined above as a necessary background to any modeling approach to unobserved confounders. However, a comprehensive review of instrumental variables falls way beyond the scopes of this introduc-

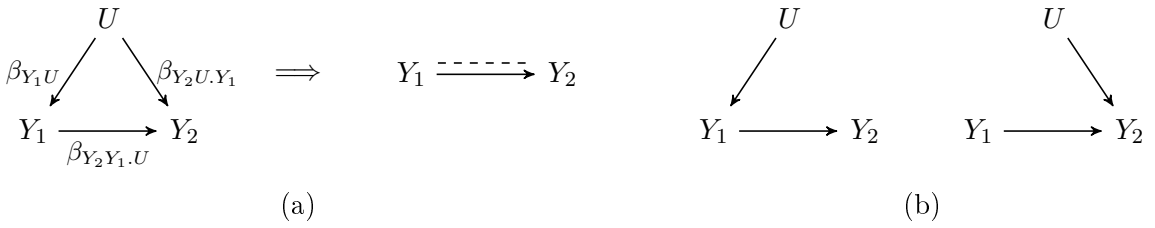

Figure 1.1: Panel (a): direct confounding effect: the variable U is explanatory to both the response of interest, Y_2, and and its predictor, Y_1. The unobservability of U induces a distorted association between Y_2 and Y_1 (dashed line) other than the one the researcher is interested to estimate, $\beta_{Y_2Y_1.U}$. Panel (b): no distortion is experienced whenever $Y_2 \perp\!\!\!\perp U|Y_1$ or $Y_1 \perp\!\!\!\perp U$, even though U is unmeasured.

tion, which is instead limited to the sole concepts relevant for the thesis.

1.3.2 Distortion of Effects Induced by Direct Unmeasured Confounding

Following Wermuth and Cox (2008), I consider the simplest instance of direct unmeasured confounding as shown in Panel (a) of Figure 1.1; the direction of the edges denotes that U is explanatory to both Y_1 and Y_2, whereas Y_1 only to Y_2. I also use U to indicate that this variable is not observed by the researcher. For simplicity, it is further assumed that the random variables have marginally zero means and variances σ_j^2's, $j \in \{Y_1, Y_2, U\}$. The corresponding data generating process is then given by the following recursive (or triangular) system of linear equations:

$$Y_1 = \beta_{Y_1U}U + \varepsilon_{Y_1}, \quad Y_2 = \beta_{Y_2Y_1.U}Y_1 + \beta_{Y_2U.Y_1}U + \varepsilon_{Y_2}, \quad U = \varepsilon_U, \tag{1.9}$$

where each stochastic component ε_j has mean zero and is uncorrelated with the explanatory variables of the right-hand side of the corresponding equation.

The unobservability of U implies that, instead of basing inference on the conditional density $f_{Y_2|Y_1,U}(y_2|y_1,u)$, one conducts the analysis on $f_{Y_2|Y_1}$, as obtained by marginalizing over U (Cox and Wermuth, 2003). Namely, from the factorisation of the joint density of $f_{Y_2,Y_1,U}$, it holds that

$$f_{Y_2|Y_1}(y_2|y_1) = \int f_{Y_2|Y_1,U} f_{U|Y_1} \mathrm{d}u.$$

However, this practice induces a distortion in the dependence of the relevant variables (i.e. $Y_1 \longrightarrow Y_2$), which consists now of the coefficient $\beta_{Y_2Y_1.U}$ and an effect due to the indirect association of Y_2 on Y_1 via U (as represented by the dashed line in the figure). By introducing the additional assumption of the Gaussianity of the error terms, it is possible to provide an

explicit form of these distorted effects. To this end, consider first

$$\mathbb{E}[Y_2|Y_1] = \beta_{Y_2Y_1.U}Y_1 + \beta_{Y_2U.Y_1}\mathbb{E}[U|Y_1] \tag{1.10}$$

and notice that the expected value of U given Y_1 can be obtained explicitly by standard results of the Gaussian distribution:

$$\mathbb{E}[U|Y_1] = \rho_{U,Y_1}(\sigma_U/\sigma_{Y_1})Y_1 = \beta_{Y_1U}(\sigma_U^2/\sigma_{Y_1}^2)Y_1,$$

where ρ_{U,Y_1} denotes the correlation coefficient between U and Y_1. Hence (1.10) can be re-written as

$$\mathbb{E}[Y_2|Y_1] = (\beta_{Y_2Y_1.U} + \beta_{Y_2U.Y_1}\beta_{Y_1U}\sigma_U^2/\sigma_{Y_1}^2)Y_1 =: \beta_{Y_2Y_1}Y_1, \tag{1.11}$$

from which it is clear that a distortion of effects occurs in general unless either $\beta_{Y_2U.Y_1} = 0$ or $\beta_{Y_1U} = 0$. These conditions are equivalent to set either $Y_2 \perp\!\!\!\perp U|Y_1$ or $Y_1 \perp\!\!\!\perp U$, whose graphical representations are given in Panel (b) of Figure 1.1 (e.g., Cox and Wermuth, 1993). It is worth mentioning that the latter condition is usually attained by study design, and is satisfied whenever a randomisation mechanism is used successfully to allocate individuals to treatment Y_1.

Within this simplified framework, it is also possible to relate nicely the distortion of effects due to the unobservability of U and the inconsistency of the parameters' estimates stressed in the econometric literature. Specifically, by writing the linear projection of U onto the observed covariate Y_1 of Y_2 as (Wooldridge, 2002, p. 62)

$$U = \beta_{UY_1}Y_1 + \varepsilon_U, \quad \mathbb{E}[\varepsilon_U|Y_1] = \mathbb{E}\varepsilon_U = 0,$$

it follows that

$$\beta_{UY_1} = \sigma_{U,Y_1}/\sigma_{Y_1}^2 = \beta_{Y_1U}(\sigma_U^2/\sigma_{Y_1}^2),$$

and

$$Y_2 = (\beta_{Y_2Y_1.U} + \beta_{Y_2U.Y_1}\beta_{Y_1U}\sigma_U^2/\sigma_{Y_1}^2)Y_1 + (\beta_{Y_2U.Y_1}\varepsilon_U + \varepsilon_{Y_2}).$$

Moreover, since

$$\mathbb{E}[\beta_{Y_2U.Y_1}\varepsilon_U + \varepsilon_{Y_2}|Y_1] = \beta_{Y_2U.Y_1}\mathbb{E}[\varepsilon_U|Y_1] + \mathbb{E}[\varepsilon_{Y_2}|Y_1] = 0,$$

it is just a matter of simple algebra and standard arguments to characterise the inconsistency of the least squares estimator (LSE) under the omission of U, $\widehat{\beta}_{Y_2Y_1}$, as

$$\widehat{\beta}_{Y_2Y_1} \xrightarrow{p} \beta_{Y_2Y_1.U} + \beta_{Y_2U.Y_1}\beta_{Y_1U}\sigma^2_U/\sigma^2_{Y_1} \neq \beta_{Y_2Y_1.U},$$

where the right-hand side is what a researcher would have estimated in case of the full observability of U. In this simplified scenario, therefore, the distortion of effects due to direct unmeasured confounding coincides with the limit in probability of the LSE.

1.3.3 IVs as Solution to the Endogeneity Problem

Previous section argued that, wherever observational studies are affected by unmeasured confounding, the simple linear regression model in general fails to provide a faithful representation of the true association between the relevant variables in the study. Although the random allocation of treatment Y_1 is a possible solution to the problem, randomisation may not be possible for many potential questions of interest. For example, if a researcher wants to assess the effect of alcohol consumption on some health outcome, the random allocation of people to different levels of alcohol consumption may rise ethical concerns as well as face legal barriers. Methods exploiting instrumental variables provide a feasible alternative solution. Loosely speaking, an IV is an observed variable that is predictive to the treatment but has no direct effect on the response and is independent of the unobserved confounders (Didelez et al., 2010).

Let Z be such a variable, and assume that the following "core" conditions hold (e.g., Didelez and Sheehan, 2007): (i) $Z \perp\!\!\!\perp U$, (ii) $Y_1 \not\perp\!\!\!\perp Z$ and (iii) $Y_2 \perp\!\!\!\perp Z|(Y_1, U)$. In particular, the latter assumption requires that the instrument and the response are independent conditionally to Y_1 and the confounder U. In other words, any IV is allowed to affect the outcome only through its dependence on Y_1. Figure 1.2 shows graphically assumptions (i)-(iii), and the corresponding generating process is now described by the system (Cox and Wermuth, 2004)

$$Y_1 = \beta_{Y_1U.Z}U + \beta_{Y_1Z.U}Z + \varepsilon_{Y_1}, \quad Y_2 = \beta_{Y_2Y_1.U}Y_1 + \beta_{Y_2U.Y_1}U + \varepsilon_{Y_2}, \quad U = \varepsilon_U, \quad Z = \varepsilon_Z.$$

As above, the aim of the analysis is to estimate consistently the association between Y_1 and Y_2, $\beta_{Y_2Y_1.U}$, under the unobservability of U. Marginalisation over the confounder yields the estimable relations $\mathbb{E}[Y_1|Z] = \beta_{Y_1Z.U}Z$ and $\mathbb{E}[Y_2|Y_1] = \beta_{Y_2Y_1}Y_1$, where the latter has been

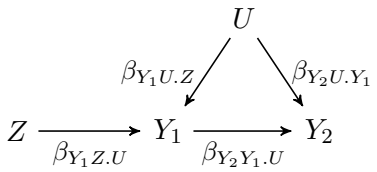

Figure 1.2: Instrumental variable Z in action: the association of interest $\beta_{Y_2Y_1.U}$ is now consistently estimable under the IV "core" conditions.

previously defined in (1.11). Hence, under core assumption (i), it follows that $\mathbb{E}[U|Z] = \mathbb{E}U = 0$ and

$$\mathbb{E}[Y_2|Z] = \beta_{Y_2Y_1.U}\mathbb{E}[Y_1|Z] + \beta_{Y_2U.Y_1}\mathbb{E}[U|Z] + \mathbb{E}[\varepsilon_{Y_2}|Z] = \beta_{Y_2Y_1.U}\beta_{Y_1Z.U}Z,$$

from which the classical IV emerges as the sample analogue of

$$\beta_{Y_2Y_1.U} = \mathbb{E}[Y_2|Z]/\mathbb{E}[Y_1|Z] = \beta_{Y_2Z}Z/\beta_{Y_1Z}Z = \sigma_{Y_2,Z}/\sigma_{Y_1,Z}, \tag{1.12}$$

namely the ratio between the two least squares regressions of Y_2 on Z and Y_1 on Z: $\widehat{\beta}_{Y_2Y_1.U} = \widehat{\sigma}_{Y_2,Z}/\widehat{\sigma}_{Y_1,Z}$. In addition, the consistency of the above estimator is guaranteed by those of $\widehat{\sigma}_{Y_2,Z}$ and $\widehat{\sigma}_{Y_1,Z}$.

Despite this appealing result, some drawbacks are commonly pointed out in the studies on instrumental variables (Cox and Wermuth, 2004). At first, marginalisation over U implies that the system for Y_1, Y_2 and Z is saturated, namely a model in which no conditional independences are present. Therefore, core conditions (i) and (iii) cannot be tested empirically but only justified on a subject-matter ground. Moreover, the estimate $\widehat{\beta}_{Y_2Y_1.U}$ turns out to have low precision unless the denominator in (1.12) is well determined. In fact, as Bound et al. (1995) showed, upon computing the inconsistency of the IV estimator relative to the LSE

$$\frac{\text{plim}_{n\to\infty}\widehat{\beta}_{Y_2Y_1.U} - \beta_{Y_2Y_1.U}}{\text{plim}_{n\to\infty}\widehat{\beta}_{Y_2Y_1} - \beta_{Y_2Y_1.U}} = \frac{\rho_{Z,\varepsilon_{Y_2}}/\rho_{Y_1,\varepsilon_{Y_2}}}{\rho_{Y_1,Z}},$$

it holds that even small departures from condition (i) are exacerbated by the poor association between the endogenous regressor and the instrument used (i.e. *weak instruments*). This issue may be particularly problematic in applied works because, as previously remarked, the special independence assumptions characterising an IV cannot be tested empirically, and the assumption $\rho_{Z,\varepsilon_{Y_2}} = 0$ may not hold exactly. Techniques for dealing with weak instruments are reviewed and discussed in Stock et al. (2002). Their paper surveys existing

hypothesis tests constructed for detecting weak instruments and how it is possible to perform robust inference under this occurrence. Moreover, the k-class estimators is discussed as an alternative "partially robust" methodology: this class, in fact, is shown to provide more reliable estimates than classical IV under weak instruments. Well-known estimators, like the Limited-information Maximum Likelihood (LIML), Fuller-k (Fuller, 1977) and Jacknife IV (Angris et al., 1999) belong all to this class.

1.4 Epilogue

This thesis builds around an extension of the approach outlined in Section 1.3 for ordinal polychotomous variables Y_1 and Y_2 with covariate effects modelled flexibly using penalized regression splines as in Section 1.2.2. To this end, a bivariate recursive system of equations along the lines of (1.9) is set up as the prototypical model of the thesis, whose representation and estimation are detailed in Chapter 2. The method discussed is then suitable to account for the presence of unmeasured confounding in the study of interest. An application concerning the relationship between education achievements and drinking behaviour of young adults in the UK is also discussed. In the economic literature, in fact, it is often argued that both the choice of education and the undertaking of healthy behaviours (including drinking and smoking) are driven by a subjective time preference discounting future well-being. Since this is hardly measurable by researchers, specific models need to be employed.

The role of unobserved variables in applied research, however, is not only confined to the confounding effect. Some situations may instead require multivariate models to estimate more efficiently the association between some variables of interest. For example, this is the case in multi-parties vehicle collisions where common variables and crash dynamics may influence simultaneously the level of severity injuries sustained by all the people involved in the same accident. However, common risk factors may not be reported in the police records (e.g., the speed at the time of the hit, road pavement conditions, etc.) and hence omitted from the analysis. Larger standard errors are then expected if each party in a vehicle accident is not modelled jointly with all the others. Chapter 3 considers this instance within a bivariate system of seemingly unrelated regression equations. Model flexibility is further enhanced by the introduction of smoothers for spatially-dependent covariates and the dependence structure between the responses represented beyond the classical bivariate Standard Normal distribution.

Models described in Chapters 2 and 3 are developed by acknowledging a common structure for multivariate GAMs with discrete responses. This can be described in terms of a generic penalized GLM framework, where different penalisation terms are employed to achieve various kinds of covariate effects, including non-parametric and spatially-dependent ones. Chapter 4 outlines this generalisation and applies it to a model for non-random sample selection. Specifically, I address the problem of individuals who select themselves into (or out of) the relevant sample because of the presence of some unobservable personal characteristics. In this case, randomness assumption is violated and any inference conducted on the given sample is therefore invalid. The employment of a bivariate system of equation helps correcting the issue. In fact, this structure models hierarchically the decision of leaving or staying into the sample as the resultant of a dichotomous choice, the distribution of which would affect that of the response of primary interest. This idea is illustrated by replicating the analysis on HIV prevalence in Zambia recently proposed in Marra et al. (2015), whose model representation is naturally nested in the described generic framework.

In conclusion, the developments contained in the thesis have been collected in the following papers (listed in chronological order):

– "On Some Fundamental Order Relations Implied by a Multivariate Cumulative Link Model", *submitted*.

  This paper is an attempt to reconcile some background assumptions defining CLMs with the language and properties of the order theory. I first define *coherent* any CLM in which the order relations in the set $\mathcal{K}$ are maintained in some sense in the codomain of the link function, g^{-1}, the real line. I subsequently provide conditions for coherency. Specifically, it is required that g is an order-embedding, a criterion of which is the monotonicity of the sequence $\{c_k\}_k$. This is trivial in the univariate case and extends intuitively also to the multivariate framework. To see this, let now $\mathcal{K}^J$ be the J-fold product of $\mathcal{K}_j$, where $(\mathcal{K}_j, \preceq)$ is totally ordered for any $j = 1, \ldots, J$. Then, provided that $\{c_{j,k_j}\}$ is an increasing sequence in k_j for any j, the set of non-overlapping hyper-rectangles in $\mathbb{R}^J$ with vertices the cut points is proved to be isomorphic to $\mathcal{K}^J$. Stated differently,

$$\{\boldsymbol{\eta}_{\bar{k}} \leq \boldsymbol{\eta}_k\} \text{ in } \mathbb{R}^J \iff \{\bar{k} \preceq k\} \text{ in } \mathcal{K}^J,$$

  with $k := (k_1, \ldots, k_J)$, $\boldsymbol{\eta}_k := (\eta_{1,k_1}, \ldots, \eta_{J,k_J})^\top$ and $\bar{k}, k \in \mathcal{K}^J$. This paper has not been included in the thesis for no good reasons, however some of its terminology and

ideas have been variously incorporated in the proceeding work. Its main results are explained in Chapter 4.

– "Discrete Responses in Bivariate Generalized Additive Models", *arXiv:1508.01302v1* (with G. Marra).

– "Semi-parametric Bivariate Polychotomous Ordinal Regression", *Statistics and Computing (in press)* (with G. Marra).

– "A Copula Additive Model for Ordinal Responses with Application to Vehicle Accident Severity Injuries", *submitted* (with G. Marra).

Chapter 2

Semi-parametric Bivariate Polychotomous Ordinal Regression

A pair of polychotomous random variables $(Y_1, Y_2)^\top =: \boldsymbol{Y}$, where each Y_j has a totally ordered support, is studied within a penalized Generalized Linear Model framework. We deal with a triangular generating process for $\boldsymbol{Y}$, a structure that has been employed in the literature to control for the presence of residual confounding in observational studies. Differently from previous works, however, the proposed model allows for a semi-parametric estimation of the covariate-response relationships. In this way, the risk of model mis-specification stemming from the imposition of fixed-order polynomial functional forms is also reduced. The proposed estimation methods and related inferential results are finally applied to study the effect of education on alcohol consumption among young adults in the UK.

Published as: Donat, F. and G. Marra (2015), Semi-parametric Bivariate Polychotomous Ordinal Regression, *Statistics and Computing (in press)*.

2.1 Introduction

Polychotomous ordinal data arise in many areas of statistical analysis and are particularly frequent in surveys and observational studies. Several questions may be asked to measure people's feelings on a matter of interest, as well as some relevant information reported on a monotonic scale. Examples include individuals' perceived social class or their educational attainments. Since it is usually acknowledged that these types of data possess levels that can be "naturally" ordered, it is desirable to account for this feature in any model's representation and estimation. Specific methodologies were developed to address this issue, starting from the seminal works of Aitchison and Silvey (1957) and Snell (1964), up to their modern expressions of the Cumulative Link Models (CLM; McCullagh, 1980) in which ordinal responses are represented in the form of Generalized Linear Models (GLMs; Nelder and Wedderburn, 1972). In general, a CLM links the cumulative distribution function of an ordinal discrete random variable to a level-specific predictor through a known link function.

However, despite this similarity with generalized linear models, CLMs are not formally a member of the class of (univariate) GLMs, rather they are shown to belong to a class of multivariate generalized linear models (Fahrmeir and Tutz, 2001). Notably, many analogies are shared by CLMs and GLMs which are made explicit and employed in the development of the chapter. An interesting historical review discussing merits (and limits) of each of the above contributions can be found in the monograph of Greene and Hensher (2010).

This chapter deals with a bivariate system of polychotomous outcomes, $\boldsymbol{Y} := (Y_1, Y_2)^\top$, where each Y_j, for $j = 1, 2$, is measured on the ordinal scale. To fix ideas, and recalling that many discrete data can be modelled as a coarse version of a continuous latent random variable Y_j^* (e.g. McKelvey and Zavoina, 1975, Anderson and Philips, 1981), we anticipate that the aim of this work is to estimate and make inference from a model with the following structure

$$\begin{array}{rcl} Y_1^* & = & \mathbf{x}_1^\top \boldsymbol{\beta}_1 + s_{1,1}(\mathrm{v}_{1,1}) + \cdots + s_{1,L_1}(\mathrm{v}_{1,L_1}) + \varepsilon_1 \\ Y_2^* & = & \psi Y_1^* + \mathbf{x}_2^\top \boldsymbol{\beta}_2 + s_{2,1}(\mathrm{v}_{2,1}) + \cdots + s_{2,L_2}(\mathrm{v}_{2,L_2}) + \varepsilon_2 \end{array}, \tag{2.1}$$

where the s_{j,l_j} are unknown smooth functions appropriately represented and fitted and $\psi \in \mathbb{R}$. Upon setting $\boldsymbol{Y}^* := (Y_1^*, Y_2^*)^\top$, $\mathbf{X} := \mathrm{diag}(\mathbf{x}_1^\top, \mathbf{x}_2^\top)$, $\boldsymbol{\beta} := \mathrm{vec}(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$ and $\boldsymbol{\varepsilon} := (\varepsilon_1, \varepsilon_2)^\top$ we re-write (2.1) in the more compact form $\boldsymbol{\Gamma Y}^* = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ or

$$\mathbf{L}\boldsymbol{Y}^* = \mathbf{L}\boldsymbol{\Gamma}^{-1}(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) \qquad \boldsymbol{\varepsilon} \sim \mathcal{N}_2(\mathbf{0}_2, \boldsymbol{\Omega}), \tag{2.2}$$

where

$$\boldsymbol{\Omega} := \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}, \quad \boldsymbol{\Gamma} := \begin{bmatrix} 1 & 0 \\ -\psi & 1 \end{bmatrix}, \quad \mathbf{L} := \begin{bmatrix} 1 & 0 \\ 0 & (\sqrt{1 + 2\psi\rho + \psi^2})^{-1} \end{bmatrix} \quad \text{and}$$

$\rho \in [-1, 1]$ is the correlation coefficient. It follows that $\mathbf{L}\boldsymbol{Y}^* \sim \mathcal{N}_2(\mathbf{L}\boldsymbol{\Gamma}^{-1}\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma} = \mathbf{L}\boldsymbol{\Gamma}^{-1}\boldsymbol{\Omega}\boldsymbol{\Gamma}^{-\top}\mathbf{L}^\top$. Model (2.2) represents the so-called reduced-form of system (2.1) which is termed triangular (or recursive) given that $\boldsymbol{\Gamma}$ is a lower triangular matrix.

Apart from a pure methodological interest, the study of (2.2) is motivated by the practical issue of analysing data affected by residual confounding. This is a situation where an unknown or not readily quantifiable variable is associated with both an ordinal response of interest and an ordinal treatment. When not adequately controlled for, unmeasured confounders may pose serious limitations to the use of standard estimators as they usually yield inconsistent estimates. An analogous bivariate system of equations for dichotomous out-

comes addressing this problem has been recently discussed by Radice et al. (2015). Their paper, however, considers only the recursion to occur only at the level of the manifest Y_1, namely they set $\psi = 0$ and augment $(Y_2|\mathbf{x}_2^\top)$ with $Y_1 \in \{0, 1\}$ accordingly. At present, the only alternatives available to model ordinal polychotomous responses in a similar (albeit purely parametric) fashion comprise the routines of Sajaia (2008) and the mixed effects version proposed by Buscha and Conte (2014), both for the `STATA` computational environment (StataCorp, 2015). The first contribution of this chapter, therefore, concerns the development of an approach for fitting system (2.1), which permits the semi-parametric estimation of the covariate-response relationships. This allows us to determine the functional form of covariate effects from the data without the imposition of finite-order polynomials, hence reducing the risk of mis-specification. Moreover, semi-parametric modelling avoids categorising continuous variables into groups based on intervals or frequencies. This approach, which is often employed in empirical studies, is not immune of disadvantages. It introduces the further issue of defining cut-points, and assumes *a priori* that the relationship between the response and the categorised covariates is flat within the chosen intervals (Royston and Altman, 1994).

In principle, once a distributional assumption for the latent random vector $\boldsymbol{Y}^*$ is made and an observational rule for the manifest polychotomous responses established, the likelihood function of the model can be easily set up and the parameters estimated. The approach we take here, however, is slightly different and more general. In line with Peyhardi et al. (2014), we specify a GLM class for bivariate ordinal responses defined by the triplet $(r, F_2, \mathbf{Z})$, where F_2 and $\mathbf{Z}$ denote a 2-variate distribution function and the design matrix, respectively, while r is a map characterising the types of response vector. We then describe the class of penalized GLMs and show that (2.2) can be specified as an instance of it. In this way, a generic algorithm for the estimation and inference of any penalized GLMs endowed with the $(r, F_2, \mathbf{Z})$ representation can be developed, and hence potentially applied to any other multivariate model for discrete responses with semi-parametric covariate effects. At a smaller scale, this is already achieved in this work: we also discuss the representations corresponding to a mixture of dichotomous and polychotomous outcomes, as well as some other models nested in the triangular structure. For instance, our framework also comprises the seemingly unrelated regression equations model (SURE) of Hillmann et al. (2014), which is recovered by setting $\mathbf{L} = \mathbf{I}_2$, and allows for the estimation of a bivariate system of correlated ordinal probit regressions.

After having represented the triangular structure in a suitable penalized GLM form, Section 2.3 is devoted to the description of the corresponding estimation algorithm. It is worth stressing that the triplet $(r, F_2, \mathbf{Z})$ is all it is needed for this scope, since it already incorporates the information concerning the model specification, link function used, and types of responses. In this way, the description of a more general model will enable us to develop an algorithm suitable for any other model belonging to the class. The approach we follow is analogous to those of Vector Generalized Additive Models (VGAM; Yee and Wild, 1996) and structured additive regressions models by Klein et al. (2015) and Klein and Kneib (2015). All the necessary computational routines are incorporated in the `R` function `SemiParCLM` that accompanies this chapter. Finally, the functioning of our model is illustrated in Section 2.5 using data from the BCS70 dataset (UCL Institute of Education. Centre for Longitudinal Studies, 2007), aiming at quantifying the effect of education on alcohol consumption among young adults in the UK.

2.2 A GLM Representation for Bivariate Ordinal Responses

Let us assume that we observe realisations from the distribution of a bivariate random vector $\boldsymbol{Y} = (Y_1, Y_2)^\top$ with discrete support $\mathcal{K} := \mathcal{K}_1 \times \mathcal{K}_2$ and such that $(\mathcal{K}_j, \preceq)$ is totally ordered for any $j \in \mathcal{J} = \{1, 2\}$. Specifically, we consider the set $\mathcal{K}_j := \{1, \ldots, k_j, \ldots, K_j\}$ with $\#(\mathcal{K}_j) = K_j < \infty$, where k_j represents a natural number. We then say that variable Y_j shows finite K_j levels. Notice that the totality assumption implies the comparability of each k_j with respect to all the remaining elements in $\mathcal{K}_j \setminus \{k_j\}$. In other words, the proposed methodology is only applicable in those situations where it is possible to state whether $\bar{k}_j \preceq k_j$ or $k_j \preceq \bar{k}_j$ for any $k_j, \bar{k}_j \in \mathcal{K}_j$. For example, this may not be the case in surveys foreseeing the possibility to tick the "don't know" box. Whenever this instance is likely to occur, more appropriate models for partially ordered responses have to be employed, like the one discussed by Zhang and Ip (2012). In particular, their approach decomposes the finite lattice describing the support of $\boldsymbol{Y}$ into chains and anti-chains and, for each of the two, employs models for ordinal and nominal responses, respectively.

Covariate information is collected in the vector $\boldsymbol{x} := \text{vec}(\boldsymbol{x}_1, \boldsymbol{x}_2)$, where $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are available regressors. It is then possible to set up a model relating the conditional probability $\pi_k := \mathbb{P}[\boldsymbol{Y} = k | \boldsymbol{X} = \boldsymbol{x}]$, with $k := (k_1, k_2) \in \mathcal{K}$, to $\boldsymbol{x}$ through the GLM form (Peyhardi

et al., 2014)

$$\bar{\boldsymbol{\pi}} = \boldsymbol{g}^{-1}(\boldsymbol{\eta}) := (\boldsymbol{r}^{-1} \circ \mathcal{F})(\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_{K-1}) \in [0,1]^{\#(\mathcal{K})-1}, \tag{2.3}$$

where $\mathcal{F}(\boldsymbol{\eta}) := (F_2(\boldsymbol{\eta}_1), \ldots, F_2(\boldsymbol{\eta}_{K-1}))^\top$, $F_2 : \mathbb{R}^2 \longrightarrow (0,1)$ denotes any fully-specified bivariate distribution function and $K := (K_1, K_2)$. A bivariate CLM for polychotomous ordinal responses is then recovered by setting $\boldsymbol{r}(\bar{\boldsymbol{\pi}}) := (r(\pi_k))_{k \in \mathcal{K} \setminus \{K\}}$ where, for each k,

$$r(\pi_k) := \mathbb{P}[Y_1 \preceq k_1, Y_2 \preceq k_2 | \boldsymbol{X} = \boldsymbol{x}] = \sum_{\tilde{k}_1 \preceq k_1} \sum_{\tilde{k}_2 \preceq k_2} \pi_{\tilde{k}_1, \tilde{k}_2}.$$

The array $\boldsymbol{\eta}_k := (\eta_{1,k_1}, \eta_{2,k_2})^\top \in \mathbb{R}^2$ defines the linear predictor of the model, and embodies the functional form of the covariate effects. Since this is pivotal in our proceeding discussion, it will be analysed more precisely in Section 2.2.1. In line with the multivariate nature of the model, the generic pair $(k_1, k_2) \in \mathcal{K}$ is assumed to follow a lexicographical order, that is $(\bar{k}_1, \bar{k}_2) \preceq (k_1, k_2)$ if and only if $\bar{k}_1 \preceq k_1$ or $(\bar{k}_1 = k_1 \ \wedge \ \bar{k}_2 \preceq k_2)$ for $\bar{k}_j, k_j \in \mathcal{K}_j$. We stress that:

Remark 1. *Any regression model for ordinal outcomes sets two constraints in representation (2.3). The obvious one requires $r(\pi_k) = r(\pi_{\bar{k}}) + \pi_k \geq r(\pi_{\bar{k}})$ for $\bar{k} := (k_1, k_2 - 1) \preceq (k_1, k_2) =: k$; whereas $\boldsymbol{\eta}_{\bar{k}} \leq \boldsymbol{\eta}_k$ is needed for all $\bar{k} \preceq k$ and $\bar{k}, k \in \mathcal{K}$. In particular, the latter can be thought of as a model coherency condition and is introduced to ensure that the order relations implied by the set $\mathcal{K}$ are maintained in the domain of the linear predictor, $\mathbb{R}^2$. These issues will be illustrated with greater details in Section 4.2.3.*

To meet these requirements, let us set a pair of cut points (or threshold parameters), collected in the vector $\boldsymbol{c}_k := \{(c_{1,k_1}, c_{2,k_2})^\top \in \mathbb{R}^2 | c_{j,\bar{k}_j} \leq c_{j,k_j}, \forall \bar{k}_j \preceq k_j, k_j \in \mathcal{K}_j, \forall j\}$, and such that $c_{j,K_j} = \infty$ and $c_{j,0} := c_{j,1-1} = -\infty$. We consequently define a bivariate probit regression for ordinal responses as

$$r(\pi_k) = \Phi_2(\boldsymbol{c}_k - \mathbf{X}\boldsymbol{\beta}) = \Phi_2(\mathbf{Z}\boldsymbol{\beta}_k), \tag{2.4}$$

where $\mathbf{Z} := \text{diag}(\mathbf{z}_1^\top, \mathbf{z}_2^\top)$ is the analogue of the design matrix, $\mathbf{z}_j := (1, -\mathrm{x}_{j,1}, \ldots, -\mathrm{x}_{j,M_j})^\top$, and $\boldsymbol{\beta}_k := \text{vec}(\boldsymbol{\beta}_{1,k_1}, \boldsymbol{\beta}_{2,k_2})$, $\boldsymbol{\beta}_{j,k_j} := (c_{j,k_j}, \beta_{j,1}, \ldots, \beta_{j,M_j})^\top \in \mathbb{R}^{M_j+1}$, is the vector of regression coefficients. M_j is used to denote the number of covariates included in equation j. Finally, the linear predictors are given by $\boldsymbol{\eta}_k := \mathbf{Z}\boldsymbol{\beta}_k$, so that GLM form (2.4) can be characterised by the triplet $(r, F_2, \mathbf{Z})$. Notice that we have set $F_2 \equiv \Phi_2$ in the proposed model specification.

The above definition of the cut points relies on the weak monotonicity assumption of

$\{c_{j,k_j}\}_{k_j}$ for all j. In fact, although Dale (1986) required this sequence to be *strictly* increasing to ensure that each π_k is positive, we regard this condition too stringent, as it eventually adds a further unnecessary constraint to the likelihood function. Admittedly, as Haberman (1980) pointed out for the univariate case, wherever two subsequent cut points are congruent (e.g., when zero counts are observed for a given level as shown in Pratt, 1981) the Maximum Likelihood Estimator (MLE) is located at the boundary of the parameter space. Notwithstanding, since the estimates so obtained are still admissible *per se* because there is no ambiguity in reporting an estimate on the boundary of the parameter space, it seems to us that the exclusion of the case $c_{j,k_j} = c_{j,k_j+1}$ is formally restrictive and thus to be avoided. Some alternative estimators to the MLE dealing with this issue have been recently proposed by Kosmidis (2014) for univariate CLMs. Specifically, his paper introduces a reduced-bias estimator of cumulative link models for ordinal data which is shown to generalise well-known constant adjustments used to regularise maximum likelihood estimates and hence offers a solution to boundary estimates.

Remark 2. *Although the focus is on the modelling of ordinal responses, our methodology is immediately applicable also to mixtures of dichotomous and polychotomous variables. To see this, let us first decompose $r = r_2 \circ r_1$, where the subscripts correspond to the elements of the 2-dimensional vector $\boldsymbol{Y}$ they refer to. Moreover, the inclusion of a binary outcome in (2.3), say $Y_{\bar{\jmath}}$, corresponds to define $r_{\bar{\jmath}}$ as $\pi_{k_{\bar{\jmath}}} \mapsto \pi_{k_{\bar{\jmath}}}$, the identity map. Then it follows*

$$(r_{\bar{\jmath}} \circ r_j)(\pi_k) = \pi_{k_{\bar{\jmath}},1} + \cdots + \pi_{k_{\bar{\jmath}},k_j}.$$

Notice that the fact we have put $k_{\bar{\jmath}}$ before k_j was just for notational convenience. In fact, it is indifferent to the order in which the different types of variables appear. More formally, since $r_{\bar{\jmath}}$ is the identity map, we have that the function composition is commutative:

$$(r_j \circ r_{\bar{\jmath}})(\pi_k) = r_j(\pi_k) = r_{\bar{\jmath}}(r_j(\pi_k)) = (r_{\bar{\jmath}} \circ r_j)(\pi_k)$$

for $j, \bar{\jmath} \in \mathcal{J}$ and every $k \in \mathcal{K}$.

In the proceeding discussion, we extend representation (2.4) to account for semi-parametric model components, and develop a generic estimation algorithm for a bivariate system of polychotomous ordinal responses expressible in the $(r, F_2, \mathbf{Z})$ form.

2.2.1 Penalized Regression Spline Representation

Each linear predictor $\boldsymbol{\eta}_k$ can be specified to embody different types of covariate effects. In this work, additive non-parametric effects of the continuous regressors v_{j,l_j} are represented using penalized regression splines (Eilers and Marx, 1996). Let us assume we observe a sample of n individuals indexed by the subscript i, and let $\{\mathrm{v}_{j,l_j,(1)},\ldots,\mathrm{v}_{j,l_j,(i)},\ldots,\mathrm{v}_{j,l_j,(n)}\}$ be the ordered vector of corresponding observations. Thus, provided that we can choose a rich enough set of basis functions, $\boldsymbol{b}_{j,l_j}$, delimited by H_j+1 knot points in the interior of $[\mathrm{v}_{j,l_j,(1)},\mathrm{v}_{j,l_j,(n)}]$, we approximate

$$s_{j,l_j}(\mathrm{v}_{j,l_j,i}) \approx \boldsymbol{\delta}_{j,l_j}^{\top}\boldsymbol{b}_{j,l_j}(\mathrm{v}_{j,l_j,i}) \in \mathbb{R}, \quad j=1,2, \;\; l_j=1,\ldots,L_j.$$

Specifically, $s_{j,l_j} : \mathbb{R} \longrightarrow \mathbb{R}$ is restricted to be a smooth function, $\boldsymbol{b}_{j,l_j}(\mathrm{v}_{j,l_j,i}) := (b_{j,l_j,h_j}(\mathrm{v}_{j,l_j,i}))_{h_j} \in \mathbb{R}^{H_j}$, and $\boldsymbol{\delta}_{j,l_j} \in \mathbb{R}^{H_j}$ is a parameter vector associated to s_{j,l_j}. Basis functions are usually chosen to have convenient mathematical properties and good numerical stability. Among the various functions supported by our implementation, the B-splines, cubic regression and thin-plate regression splines are the most widely used in applications (e.g. Ruppert et al., 2003; Wood, 2003). To achieve functions' identification, the centering constraint $\mathbf{1}_n^{\top}\boldsymbol{s}_{j,l_j}=0$ is imposed, where $\boldsymbol{s}_{j,l_j}$ denotes the vector whose i-th element is $s_{j,l_j}(\mathrm{v}_{j,l_j,i})$. This approach is incorporated automatically in our model estimation through the parsimonious method outlined in Wood (2006).

To recover a more compact and comprehensive representation of the linear predictors, we set $\boldsymbol{\beta}_{[j,l_j]} := \boldsymbol{\delta}_{j,l_j}$ which represents the sub-vector of $\boldsymbol{\beta}_j$ referring to the (j,l_j)-th smooth and, accordingly, $\mathbf{X}_{[j,l_j]} \in \mathbb{R}^{n\times H_j}$ is the matrix whose i-th row is given by $\boldsymbol{b}_{j,l_j}^{\top}(\mathrm{v}_{j,l_j,i})$. Then, we can write the linear predictor of the j-th response as

$$\boldsymbol{\eta}_j = \boldsymbol{c}_j - \mathbf{X}_{j,1}\boldsymbol{\beta}_{j,1} - \cdots - \mathbf{X}_{j,M_j}\boldsymbol{\beta}_{j,M_j} = \mathbf{Z}_j\boldsymbol{\beta}_j \in \mathbb{R}^n,$$

where $\mathbf{Z}_j := (\mathbb{I}_j, -\mathbf{X}_{j,1},\ldots,-\mathbf{X}_{j,m_j},\ldots,-\mathbf{X}_{j,M_j})$, $\boldsymbol{\beta}_j := \mathrm{vec}(\boldsymbol{c}_{j,k},\boldsymbol{\beta}_{j,1},\ldots,\boldsymbol{\beta}_{j,M_j})$, $\boldsymbol{c}_j := (c_{j,k_j,i})_i \in \mathbb{R}^n$, $\mathbb{I}_j := \mathrm{diag}(\mathbb{1}_{\mathrm{y}_{j,i}=k_j})_{j,k_j} \in \{0,1\}^{n\times K_j-1}$ and $\boldsymbol{c}_{j,k_j} := (c_{j,1},\ldots,c_{j,K_j-1})^{\top}$. Notice that each $\eta_{j,i} \in \boldsymbol{\eta}_j$ depends on a specific level k_j induced by $c_{j,k_j,i}$. Therefore, the i-th row of $(\boldsymbol{\eta}_1,\boldsymbol{\eta}_2) \in \mathbb{R}^{n\times 2}$, $\boldsymbol{\eta}_{k,i} := (\eta_{1,k_1,i},\eta_{2,k_2,i})^{\top}$, can be thought as the sampling analogous of $\boldsymbol{\eta}_k$ employed in the preceding discussion. So re-stated, the linear predictors can be employed to incorporate both non- and purely parametric covariate effects. A modelling approach of this kind is termed semi-parametric in the statistical literature.

2.2.2 The Triangular Ordered Probit Model

The previous sections have described a generic model for a bivariate ordinal polychotomous random vector. In what follows, we qualify the structure of the triangular model of interest.

Motivation Residual confounding is a relatively frequent issue in observational studies. It occurs whenever the association between a response and one (or more) of its relevant regressor(s) is distorted by the presence of an unobserved third variable which affects simultaneously the two. Such covariates are termed endogenous in the econometric literature. A researcher would be particularly interested in controlling for pertinent unmeasured confounders as they usually lead to inconsistent estimates for the whole parameter vector. In experimental studies, one possible solution is the assignment of the relevant treatment to individuals via a randomisation mechanism, whose functioning is independent of any other factor (e.g. Frosini, 2006). However, this may not be feasible in situations where the experiment design would raise ethical or legal issues, as it is frequently the case in observational studies. Models dealing with this problem have been proposed in the literature. Cox and Wermuth (2004) and Wermuth and Cox (2008), for example, described the direct confounding effect by means of graphical models. In this setting, they quantified the distortion from endogenous covariate effects under the regular assumptions of continuous responses and a generating process represented by a triangular system of equations.

In line with the bivariate recursive model introduced by Heckman (1978) for binary outcomes, we consider the instance of an endogenous variable Y_1 that is assumed to have an impact on the response of interest Y_2. Each of them is defined on the discrete and totally ordered support $\mathcal{K}_j$, with $\#(\mathcal{K}_j) \geq 2$. For example, in the next empirical study, we argue that individuals' education attainments are potentially endogenous in explaining their weekly alcohol intake, because both are affected by a common subjective attitude. This underlying variable is recognised to be time preference in the relevant economic literature, here a bivariate system of equations is employed to describe this situation. The corresponding generating process – expressed in terms of the latent variable formulation – is the one previously given in (2.1) and (2.2). Notice that, in addition to the usual distributional assumptions (e.g. Greene and Hensher, 2010), a further condition in the form of an exclusion restriction has to be imposed in the model to achieve identification (Sajaia, 2008, Buscha and Conte, 2014). This allows us to qualify the dependence of Y_1 with a relevant variable which is independent of (i) $Y_2|Y_1$, and (ii) the unmeasured confounder. We argue, for example, that the British

Ability Scale score possesses these characteristics in the real data illustration of Section 2.4.

Model Representation As most models for discrete data, ordinal polychotomous variables can also be motivated by means of a generating latent and continuous random vector, $\boldsymbol{Y}^*$, with support the extended real plane, through the equivalence (McKelvey and Zavoina, 1975)

$$\{\boldsymbol{Y} = (k_1, k_2) \subseteq \mathcal{K}\} \Longleftrightarrow \{\boldsymbol{Y}^* \in [c_{1,k_1-1}, c_{1,k_1}] \times [c_{2,k_2-1}, c_{2,k_2}] \subseteq \mathbb{R}^2\},$$

where the Cartesian product defines the non-overlapping rectangles in $\mathbb{R}^2$ whose vertices are the cut points. Using (2.2), and by noticing that the symmetric matrix $\mathbf{L}$ is positive-definite, since $1 + 2\psi\rho + \psi^2 = (1 - \rho^2) + (\rho + \psi)^2 > 0$ and its determinant positive, it holds that

$$\begin{aligned} \{\boldsymbol{Y} \preceq k\} &\Longleftrightarrow \{\boldsymbol{\Gamma Y}^* \leq \boldsymbol{c}_k\} \Longleftrightarrow \{\mathbf{L}\boldsymbol{Y}^* \leq \mathbf{L}\boldsymbol{\Gamma}^{-1}\boldsymbol{c}_k\} \\ &\Longleftrightarrow \{\mathbf{L}\boldsymbol{\Gamma}^{-1}\boldsymbol{\varepsilon} \leq \mathbf{L}\boldsymbol{\Gamma}^{-1}(\boldsymbol{c}_k - \mathbf{X}\boldsymbol{\beta})\}, \end{aligned}$$

where the equivalence is established under coherency. Hence, given the assumed Standard Normal distribution of the stochastic model components, the proposed triangular structure for a sample of size n corresponds to the setting of

$$r(\boldsymbol{\pi}) = \Phi_2\left(\mathbf{Z}\boldsymbol{\beta}(\mathbf{L}\boldsymbol{\Gamma}^{-1})^\top; \boldsymbol{\Sigma}\right) \in [0,1]^n \qquad \boldsymbol{\Sigma} = \mathbf{I}_n \otimes \mathbf{L}\boldsymbol{\Gamma}^{-1}\boldsymbol{\Omega}\boldsymbol{\Gamma}^{-\top}\mathbf{L}^\top, \tag{2.5}$$

where $\boldsymbol{\pi} := (\pi_1, \ldots, \pi_n)^\top$, $\pi_i := \mathbb{P}[\mathrm{y}_{1,i} = k_1, \mathrm{y}_{2,i} = k_2 | \mathbf{X}]$, $\mathbf{Z} := (\mathbf{Z}_1 | \mathbf{Z}_2)$ and $\boldsymbol{\beta} := \mathrm{diag}(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) \in \mathbb{R}^{M \times 2}$, with $M := M_1 + M_2$. Notice that the assumed recursive structure implies a predictor of the form $\boldsymbol{\eta} := \mathbf{Z}\boldsymbol{\beta}(\mathbf{L}\boldsymbol{\Gamma}^{-1})^\top \in \mathbb{R}^{n \times 2}$ which differs from the generic representation previously given in (2.4). Furthermore, since the quantity $\mathbf{L}\boldsymbol{\Gamma}^{-1}$ involves a non-linear combination of the elements of the p-dimensional vector $\boldsymbol{\vartheta} := \mathrm{vec}(\boldsymbol{c}_1, \boldsymbol{c}_2, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \rho) \in \mathbb{R}^{p-1} \times [-1, 1]$, it follows that $\boldsymbol{\eta}$ is non-linear in the parameter vector. Therefore, strictly speaking, the term "linear predictor" does not apply for the proposed triangular structure, and one needs to be careful in exploiting the GLM properties of this model. As we can see in the next section, some extra terms in the expressions for the score and Hessian have to be accounted for.

Finally, all the relevant model specifications nested in (2.5) are summarised in Table 2.1, in which the corresponding $(r, F_2, \mathbf{Z})$ forms are detailed. Estimation can hence proceed by employing a generic algorithm as detailed in the next section. In particular, the seemingly unrelated regression (SUR) representation is recovered by setting $\psi = 0$. This form is usually

Model	$r(\pi_k)$	$F_2(\boldsymbol{\eta}_k)$	$\boldsymbol{\eta}_k(\mathbf{Z})$
Triangular	$\sum_{\tilde{k}_1 \leq k_1} \sum_{\tilde{k}_2 \leq k_2} \pi_{\tilde{k}_1,\tilde{k}_2}$	$\Phi_2(\eta_{1,k_1}, \eta_{2,k_2}; \boldsymbol{\Sigma})$	$\mathbf{L}\boldsymbol{\Gamma}^{-1}\boldsymbol{\eta}_k$
SURE	$\sum_{\tilde{k}_1 \leq k_1} \sum_{\tilde{k}_2 \leq k_2} \pi_{\tilde{k}_1,\tilde{k}_2}$	$\Phi_2(\eta_{1,k_1}, \eta_{2,k_2}; \boldsymbol{\Omega})$	$(\eta_{1,k_1}, \eta_{2,k_2})^\top$
Independent	$\sum_{\tilde{k}_1 \leq k_1} \sum_{\tilde{k}_2 \leq k_2} \pi_{\tilde{k}_1,\tilde{k}_2}$	$\Phi(\eta_{1,k_1})\Phi(\eta_{2,k_2})$	$(\eta_{1,k_1}, \eta_{2,k_2})^\top$
$\mathcal{K}_1 = \{0,1\}$	$\pi_{k_1,1} + \cdots + \pi_{k_1,k_2}$	$\Phi_2(\eta_{1,k_1}, \eta_{2,k_2}; \boldsymbol{\Sigma})$	$\mathbf{L}\boldsymbol{\Gamma}^{-1}((-1)^{\mathbb{1}_{k_1=0}}\eta_{1,k_1}, \eta_{2,k_2})^\top$
$\mathcal{K}_2 = \{0,1\}$	$\pi_{1,k_2} + \cdots + \pi_{k_1,k_2}$	$\Phi_2(\eta_{1,k_1}, \eta_{2,k_2}; \boldsymbol{\Sigma})$	$\mathbf{L}\boldsymbol{\Gamma}^{-1}(\eta_{1,k_1}, (-1)^{\mathbb{1}_{k_2=0}}\eta_{2,k_2})^\top$

Table 2.1: $(r, F_2, \mathbf{Z})$ characterisation corresponding to structure (2.5) under different model specifications. The SUR equations set $\psi = 0$, hence $\boldsymbol{\Gamma} = \mathbf{L} = \mathbf{I}_2$ and $\boldsymbol{\Sigma} := \mathbf{L}\boldsymbol{\Gamma}^{-1}\boldsymbol{\Omega}\boldsymbol{\Gamma}^{-\top}\mathbf{L}^\top = \boldsymbol{\Omega}$. Two independent ordinal probit models are recovered by letting $\psi = \rho = 0$ so that $\boldsymbol{\Sigma} = \mathbf{I}_2$. The last two rows report the representation corresponding to mixtures of dichotomous and polychotomous responses in the triangular model as stated in Remark 2. Notice that, since only $K_j - 1$ cut points are effectively estimated, the condition $c_{j,0} := 0$ is usually set for the equation corresponding to the binary response, and the intercept is now estimable. The label $\boldsymbol{\eta}_k \in \mathbb{R}^2$ has been used to denote the i-th row of $\boldsymbol{\eta}$, which in turn depends on the level $k \in \mathcal{K}$.

employed for the joint modelling of inter-related outcomes or symmetry in the responses. This is the case, for instance, of the estimation of the injuries sustained by two people in the same car accident (Yamamoto and Shankar, 2004), or the intensity of a certain disease in humans' left and right eyes (Kim, 1995).

2.3 Estimation Methods and Inference

In this chapter, the random vector $\boldsymbol{Y}|\boldsymbol{X}$ is assumed to follow a Categorical distribution, which is a member of the exponential family of distributions. Using a random sample of conditionally independent responses given the regressors, we write the log-likelihood function for generic model (2.3) as

$$\ell(\boldsymbol{\vartheta}|\mathbf{y}_1, \mathbf{y}_2, \mathbf{X}_1, \mathbf{X}_2) = \sum_{i=1}^{n} \sum_{k \in \mathcal{K}} \mathbb{1}_{\mathrm{y}_{1,i}=k_1} \mathbb{1}_{\mathrm{y}_{2,i}=k_2} \log \pi_k(\mathbf{x}_{1,i}, \mathbf{x}_{2,i}),$$

where $\mathbf{x}_{j,i}^\top$ is the i-th row of matrix $\mathbf{X}_j$. Notice that the above expression remains valid irrespectively of the model actually used. In fact, it is the computation of each π_k that depends on the specific $(r, F_2, \mathbf{Z})$ form (these are all detailed in Table 2.1). For a bivariate polychotomous ordinal regression we have

$$\pi_k = r^{-1}(r(\pi_k)) = r(\pi_{k_1-1,k_2-1}) - r(\pi_{k_1-1,k_2}) - r(\pi_{k_1,k_2-1}) + r(\pi_{k_1,k_2}),$$

where each addendum can be computed as an instance of (2.5) for the triangular model. For every i, we set

$$\begin{aligned}\boldsymbol{\eta}_k &:= (\eta_{k_1-1,k_2-1}, \eta_{k_1-1,k_2}, \eta_{k_1,k_2-1}, \eta_{k_1,k_2}, \rho)^\top \in \mathbb{R}^4 \times [-1,1], \\ \boldsymbol{r}_k &:= (r(\pi_{k_1-1,k_2-1}), r(\pi_{k_1-1,k_2}), r(\pi_{k_1,k_2-1}), r(\pi_{k_1,k_2}))^\top \in [0,1]^4,\end{aligned}$$

so that the analytical expressions for score and Hessian are computed as

$$\nabla_{\boldsymbol{\vartheta}} \ell_i(\boldsymbol{\vartheta}) = \frac{\partial \boldsymbol{\eta}_k}{\partial \boldsymbol{\vartheta}} \left(\frac{1}{\pi_k} \frac{\partial \mathcal{F}_k}{\partial \boldsymbol{\eta}_k} \frac{\partial \pi_k}{\partial \boldsymbol{r}_k} \right) = \mathbf{D}_i^\top \mathbf{u}_i =: \boldsymbol{g}_i \tag{2.6}$$

and

$$\nabla_{\boldsymbol{\vartheta}\boldsymbol{\vartheta}^\top} \ell_i(\boldsymbol{\vartheta}) = \mathbf{D}_i^\top \left(\frac{1}{\pi_k} \frac{\partial^2 \mathcal{F}_k}{\partial \boldsymbol{\eta}_k \partial \boldsymbol{\eta}_k^\top} - \mathbf{u}_i \mathbf{u}_i^\top \right) \mathbf{D}_i + \frac{\partial^2 \boldsymbol{\eta}_k}{\partial \boldsymbol{\vartheta} \partial \boldsymbol{\vartheta}^\top} \mathbf{u}_i = \mathbf{D}_i^\top \mathbf{W}_i \mathbf{D}_i + \mathbf{K}_i. \tag{2.7}$$

Please refer to Appendix B.1 for a detailed account of the above expressions. Notice that, wherever *linear* predictors are used in the model (i.e. $\psi = 0$), $\mathbf{K}_i$ is structurally equal to $\mathbf{0}_p$, and $\mathbf{D}_i$ reduces to the usual design matrix. By appropriately extending the approach of Yee and Wild (1996), we define the arrays $\mathbf{W} := -\mathrm{diag}(\mathbf{W}_1, \ldots, \mathbf{W}_n, \mathbf{K})$, $\mathbf{D} := (\mathbf{D}_1^\top | \cdots | \mathbf{D}_n^\top | \mathbf{I}_p)^\top$ and $\mathbf{u} := \mathrm{vec}(\mathbf{u}_1, \ldots, \mathbf{u}_n, \mathbf{0}_p)$, where $\mathbf{K} := \sum_i \mathbf{K}_i$. These quantities are conveniently constructed to give global expressions for the score and Hessian analogous to those of univariate GLMs.

Our model specification also imposes two constraints on the parameters. Correlation coefficient ρ is by definition bounded in the closed interval $[-1, 1]$, whereas the threshold parameters are restricted to be a monotonic series under Remark 1. To this end, we make use of some transformations commonly employed in the literature. Specifically, ρ is set to its inverse hyperbolic tangent, namely $\tilde{\rho} := \tanh^{-1}(\rho) \in \mathbb{R}$, whilst the cut points are defined via a squared polynomial as in Sajaia (2008). That is, we have $\tilde{c}_{j,1} = c_{j,1}$ and $\tilde{c}_{j,k_j} := \sqrt{c_{j,k_j} - c_{j,k_j-1}}$ for any $k_j \in \mathcal{K}_j \setminus \{1\}$ and all j, resulting in $c_{j,k_j} = c_{j,k_j-1} + \tilde{c}_{j,k_j}^2 \geq c_{j,k_j-1}$. In line with the discussion of Section 2.2, we are effectively allowing two subsequent cut points to be congruent wherever $\tilde{c}_{j,k_j}$ is estimated as 0. To avoid clutter in the notation, in what proceeds we do not distinguish between the parameter vector $\boldsymbol{\vartheta}$ and its transformation $\tilde{\boldsymbol{\vartheta}} \in \mathbb{R}^p$, where the latter includes the quantities $\tilde{c}_{j,k_j}$ and $\tilde{\rho}$. Estimation is nonetheless intended to be over $\tilde{\boldsymbol{\vartheta}}$: that is we seek to maximise $\ell(\boldsymbol{\vartheta}|\cdot)$ with respect to $\tilde{\boldsymbol{\vartheta}}$.

2.3.1 Penalized GLM Form

Classic MLE is not suitable in semi-parametric regression. In fact, the intuitive optimisation of the model log-likelihood may give rise to over-fitted curves if smoothness is not adequately calibrated. To avoid this issue, we introduce in fitting a ridge-type penalty, namely $\mathcal{P}_{j,l_j} := \lambda_{j,l_j}\boldsymbol{\beta}_{j,l_j}^\top \overline{\mathbf{S}}_{j,l_j}\boldsymbol{\beta}_{j,l_j}$, whose role is to enforce certain properties of the (j,l_j)-th covariate. The tuning parameters $\lambda_{j,l_j} \in [0,\infty)$ govern the trade-off between smoothness and fit. At one extreme, $\lambda_{j,l_j} = 0$ assigns no penalty to the regression coefficients $\boldsymbol{\beta}_{j,l_j}$ and the corresponding estimated effect may interpolate the data points. At the other, $\lambda_{j,l_j} \to \infty$ results in the estimation of a straight line, a situation where the smoothness is maximal. The smoothing parameters are thus of paramount importance in any regression spline modelling, and need to be estimated reliably within the system.

The proposed representation is flexible enough to accommodate both purely parametric and non-parametric effects of the (j,l_j)-th covariate, where the former is achieved by setting $\overline{\mathbf{S}}_{j,l_j} = \mathbf{0}$. For non-parametric curve fitting one can specify the symmetric and positive semi-definite penalty matrix as

$$\overline{\mathbf{S}}_{j,l_j} := \int_{V_{j,l_j}} \boldsymbol{b}''_{j,l_j}(\boldsymbol{b}''_{j,l_j})^\top \mathrm{d}v_{j,l_j},$$

a measure of the curvature of the estimated (j,l_j)-th function. Introductions to this roughness penalty approach to curve estimation are given in Green and Silverman (1994) and Wood (2006), to which we refer the reader for details. Finally, after having regularised each penalty matrix to account for the centering constraint of Section 2.2.1, one can explicitly construct an overall penalisation term for the whole system as $\mathcal{P}_{\boldsymbol{\lambda}} := \boldsymbol{\vartheta}^\top \mathbf{S}_{\boldsymbol{\lambda}} \boldsymbol{\vartheta}$, where $\mathbf{S}_{\boldsymbol{\lambda}}$ corresponds to $\overline{\mathbf{S}}_{\boldsymbol{\lambda}}$ padded with zeros so that $\boldsymbol{\vartheta}^\top \mathbf{S}_{\boldsymbol{\lambda}} \boldsymbol{\vartheta} = \boldsymbol{\beta}^\top \overline{\mathbf{S}}_{\boldsymbol{\lambda}} \boldsymbol{\beta}$, with $\overline{\mathbf{S}}_{\boldsymbol{\lambda}} := \mathrm{diag}(\overline{\mathbf{S}}_{j,l_j})_{l_j,j}$.

2.3.2 Estimation Given the Smoothing Parameters

Parameter estimation is achieved by alternating two steps in the spirit of the outer iteration algorithm of O'Sullivan et al. (1986). They comprise: (i) the computation of $\boldsymbol{\vartheta}^{[\alpha+1]}$ given any fixed $\boldsymbol{\lambda}^{[\alpha]}$, and (ii) the employment of this estimate to update $\boldsymbol{\lambda}^{[\alpha+1]}$. At convergence, the resulting Maximum Penalized Likelihood Estimator (MPLE) is then

$$\widehat{\boldsymbol{\vartheta}} := \arg\max_{\boldsymbol{\vartheta}} \ell_{\mathrm{p}}(\boldsymbol{\vartheta}, \boldsymbol{\lambda}|\cdot) = \arg\max_{\boldsymbol{\vartheta}} \left\{ \ell(\boldsymbol{\vartheta}|\cdot) - \frac{1}{2}\boldsymbol{\vartheta}^\top \mathbf{S}_{\boldsymbol{\lambda}} \boldsymbol{\vartheta} \right\}. \tag{2.8}$$

Notice that the included quadratic form $\mathcal{P}_{\boldsymbol{\lambda}} = \boldsymbol{\vartheta}^\top \mathbf{S}_{\boldsymbol{\lambda}} \boldsymbol{\vartheta}$ is positive-semidefinite, and that the un-penalized log-likelihood function does not depend on the smoothing parameters. Hence, the joint estimation of $(\boldsymbol{\vartheta}, \boldsymbol{\lambda})$ through the optimisation of (2.8) would clearly result in over-fitted curves, as the optimal value of $\ell_{\mathrm{p}}(\boldsymbol{\vartheta}, \boldsymbol{\lambda}|\cdot)$ would be reached when $\widehat{\boldsymbol{\lambda}} = \mathbf{0}$.

Although in principle the MPLE can be implemented using any numerical optimisation procedure, works on bivariate discrete response modelling emphasise that considerable gains in precision and computational speed can be achieved by employing a trust-region algorithm (e.g. Marra and Radice, 2013 and Radice et al., 2015). In particular, the $[\alpha]$-th iteration of the routine solves the sub-problem

$$\begin{aligned} \min_{\boldsymbol{p}} \ \widetilde{\ell}_{\mathrm{p}} &= -\left[\ell_{\mathrm{p}}(\boldsymbol{\vartheta}^{[\alpha]}) + \boldsymbol{p}^\top \nabla_{\boldsymbol{\vartheta}^{[\alpha]}} \ell_{\mathrm{p}}(\boldsymbol{\vartheta}^{[\alpha]}) + \tfrac{1}{2} \boldsymbol{p}^\top \nabla_{\boldsymbol{\vartheta}^{[\alpha]} \boldsymbol{\vartheta}^{[\alpha]\top}} \ell_{\mathrm{p}}(\boldsymbol{\vartheta}^{[\alpha]}) \boldsymbol{p}\right]\Big|_{\boldsymbol{\lambda}=\boldsymbol{\lambda}^{[\alpha]}} \\ &\quad \text{subject to } \|\boldsymbol{p}\| \leq \Delta^{[\alpha]} \\ \boldsymbol{\vartheta}^{[\alpha+1]} &= \boldsymbol{\vartheta}^{[\alpha]} + \boldsymbol{p}^{[\alpha+1]}, \end{aligned} \tag{2.9}$$

where $\boldsymbol{p}^{[\alpha+1]} := \arg\min_{\boldsymbol{p}} \widetilde{\ell}_{\mathrm{p}}(\boldsymbol{\vartheta}^{[\alpha]}|\boldsymbol{\lambda}^{[\alpha]}, \cdot)$. The first line of (2.9) uses a quadratic approximation of the negative log-likelihood about $\boldsymbol{\vartheta}^{[\alpha]}$ (the so-called model function) in order to choose the best step $\boldsymbol{p}^{[\alpha+1]}$ within the ball centered in $\boldsymbol{\vartheta}^{[\alpha]}$ of radius $\Delta^{[\alpha]}$, the trust-region. This step is made particularly precise and quick by using the analytical score and Hessian as computed via (2.6) and (2.7).

Trust-region algorithms have a number of advantages for the proposed bivariate system of equations. Recall from Section 2.2.2 that model estimation requires the imposition of an exclusion restriction to achieve identification. In fact, although it can be argued that identification can also be obtained by functional form, in this case the log-likelihood may happen to be nearly flat in a non-negligible area around the optimum (e.g. Keane, 1992). This is also the case whenever the excluded covariate is a weak predictor of Y_1. In line-search optimisers, if a given iteration falls in such long plateau regions, the search for a subsequent step, say $\boldsymbol{\vartheta}^{[\alpha+1]}$, can occur far away from the current location $\boldsymbol{\vartheta}^{[\alpha]}$. Nonetheless, the algorithm can also locate that iteration close to $\boldsymbol{\vartheta}^{[\alpha]}$, and only marginal gains in the objective function are obtained in this case. It is also possible that the search happens so far away from $\boldsymbol{\vartheta}^{[\alpha]}$ that the evaluation of (2.8) is indefinite or not finite. Most algorithms may fail in this case, and user's intervention often required.

Trust-region methods, on the other hand, always solve sub-problem (2.9) before evaluating the objective function. Specifically, wherever this is not finite at the proposed $\boldsymbol{\vartheta}^{[\alpha+1]}$,

the step $\boldsymbol{p}^{[\alpha+1]}$ is rejected, the trust-region shrunken, and the optimisation computed again. The radius is also reduced if there is not agreement between the model and objective functions, that is in case the proposed point in the region is not better than the current one. Reversibly, if such agreement occurs, it is safe to expand the trust region for the next iteration. In summary, $\boldsymbol{\vartheta}^{[\alpha+1]}$ is accepted if it improves on $\boldsymbol{\vartheta}^{[\alpha]}$ and it does not cause problems in the evaluation of $\ell_{\mathrm{p}}(\boldsymbol{\vartheta}^{[\alpha+1]}|\boldsymbol{\lambda}^{[\alpha]})$, whereas the reduction/expansion of $\Delta^{[\alpha+1]}$ is based on the similarity between model and objective functions. This is represented schematically in Algorithm 1. A theoretical background and a general treatment of the algorithm is found in Nocedal and Wright (2006), whereas technical details on the implementation we have followed are given in Geyer (2013). The latter also discusses the necessary modifications to the sub-problem (2.9) and the radius for ill-scaled variables. It is worthwhile to remark that the discussion in the next section requires some iterations of the optimisation routine to be either of Newton-Raphson or of Fisher scoring-type. Close to the converged solution, the trust-region usually behaves like a classic unconstrained optimisation algorithm (Geyer, 2013; Nocedal and Wright, 2006), and this issue is therefore typically overcome.

Starting values for algorithm initialisation are conveniently fixed at convergence of the corresponding purely parametric version of the model. This practice is efficient and accounts for the presence of unmeasured confounding (which induces parameters' inconsistency), and hence allows us to locate starting values in a region that is reasonably close to the MPLE.

2.3.3 Smoothness Selection

Once an optimal value for $\ell_{\mathrm{p}}(\boldsymbol{\vartheta}^{[\alpha+1]}|\boldsymbol{\lambda}^{[\alpha]},\cdot)$ has been obtained by the scheme detailed above, we need to employ a proper estimator for $\boldsymbol{\lambda}^{[\alpha+1]}$. A number of different techniques have been proposed in the literature to estimate smoothing parameters in an automatic way. Among them, the Un-Biased Risk Estimator (UBRE) and the Generalized Cross Validation criterion (GCV, Craven and Wahba, 1979) share a primer position in applied research. In fact, their practical implementation is strengthened by the stable and efficient computational routines introduced by Wood (2004) in the context of GAMs. These have been made applicable and been directly incorporated in our algorithm. In particular, we adapt to the present context the UBRE criterion as the default option for its interpretation in terms of the log-likelihood Akaike Information Criterion (AIC).

Let $\mathbb{R}^{5n} \ni \mathbf{z} := \mathbf{D}\boldsymbol{\vartheta} + \overline{\mathbf{W}}^{-1}\mathbf{u}$ be the pseudo-data vector associated with the unpenalized model, as based on the Fisher Information matrix $\mathcal{I}(\boldsymbol{\vartheta}) := -\mathbb{E}[\nabla_{\boldsymbol{\vartheta}\boldsymbol{\vartheta}^\top}\ell(\boldsymbol{\vartheta})] =$

$-\mathbf{D}^\top\overline{\mathbf{W}}\mathbf{D}$, where $\mathbf{W} = \overline{\mathbf{W}} + o_p(1)$ in the large sample approximation. It holds that $\overline{\mathbf{W}} := \mathrm{diag}(\overline{\mathbf{W}}_1, \ldots, \overline{\mathbf{W}}_n, \mathbf{0}_p)$ because $\mathbf{K} = o_p(1)$. We next proceed, in analogy to GLMs, to the derivation of the corresponding penalized iteratively re-weighted least square (P-IRLS) algorithm (Green, 1984).

Assume that in the vicinity of the solution of (2.9) the corresponding step behaves like an unconstrained one, that $\mathbf{D}$ is of full rank p and $\overline{\mathbf{W}}$ is positive-definite throughout the parameter space. Then, from a quadratic approximation of $\ell_{\mathrm{p}}(\boldsymbol{\vartheta}, \boldsymbol{\lambda}|\cdot)$ about $\boldsymbol{\vartheta}^{[\alpha+1]}$ we obtain, as unique solution of the resulting non-singular $p \times p$ system of equations for $\boldsymbol{\vartheta}^{[\alpha+1]}$,

$$\begin{aligned} \boldsymbol{\vartheta}^{[\alpha+1]} &= \boldsymbol{\vartheta}^{[\alpha]} + (\mathcal{I}^{[\alpha]} + \mathbf{S}_{\boldsymbol{\lambda}}|_{\boldsymbol{\lambda}=\boldsymbol{\lambda}^{[\alpha]}})^{-1}(\mathbf{S}_{\boldsymbol{\lambda}}|_{\boldsymbol{\lambda}=\boldsymbol{\lambda}^{[\alpha]}}\boldsymbol{\vartheta}^{[\alpha]} - \boldsymbol{g}^{[\alpha]}) \\ \boldsymbol{\vartheta}^{*} &= (\mathbf{D}^\top\overline{\mathbf{W}}\mathbf{D} + \mathbf{S}_{\boldsymbol{\lambda}})^{-1}\mathbf{D}^\top\overline{\mathbf{W}}\mathbf{z}. \end{aligned}$$

For notational convenience, we have labelled $\boldsymbol{\vartheta}^* := \boldsymbol{\vartheta}^{[\alpha+1]}$ and ignored the superscript $[\alpha]$ in all the other quantities. Remarkably, these expressions involve arrays from the un-penalized log-likelihood, so that the only dependence on the smoothing parameters is through $\mathbf{S}_{\boldsymbol{\lambda}}$. So re-written, we observe that $\boldsymbol{\vartheta}^*$ is the solution of a Generalized Least Squares (GLS) normal equations problem, that is

$$\boldsymbol{\vartheta}^* = \arg\min_{\boldsymbol{t}} \left\|\overline{\mathbf{W}}^{1/2}(\mathbf{z} - \mathbf{D}\boldsymbol{t})\right\|^2 + \boldsymbol{t}^\top\mathbf{S}_{\boldsymbol{\lambda}}\boldsymbol{t} \tag{2.10}$$

for any given value of $\boldsymbol{\lambda}$. In particular, $\overline{\mathbf{W}}^{1/2}$ comes from the spectral decomposition of $\overline{\mathbf{W}}$, whose computation is fostered by its construction as a block diagonal matrix. In other words, at each iteration the estimating algorithm solves a linear regression of $\mathbf{z}$ onto the columns of $\mathbf{D}$ with weight matrix $\overline{\mathbf{W}}$ and ridge penalisation $\boldsymbol{\vartheta}^\top\mathbf{S}_{\boldsymbol{\lambda}}\boldsymbol{\vartheta}$.

With this equivalence at hand, define now $\widehat{\boldsymbol{\mu}} := \overline{\mathbf{W}}^{1/2}\mathbf{D}\boldsymbol{\vartheta}^* = \mathbf{P}_{\boldsymbol{\lambda}}\overline{\mathbf{W}}^{1/2}\mathbf{z}$ to be the plug-in estimator of the mean of $\overline{\mathbf{W}}^{1/2}\mathbf{z}$ evaluated at (2.10), and let $\mathbf{P}_{\boldsymbol{\lambda}}$ be the influence matrix

$$\mathbf{P}_{\boldsymbol{\lambda}} = \overline{\mathbf{W}}^{1/2}\mathbf{D}(\mathbf{D}^\top\overline{\mathbf{W}}\mathbf{D} + \mathbf{S}_{\boldsymbol{\lambda}})^{-1}\mathbf{D}^\top\overline{\mathbf{W}}^{1/2}.$$

Hence we propose to select $\boldsymbol{\lambda}$ through the minimisation of the expected discrepancy between

the true and the fitted curves:

$$\begin{aligned}\widetilde{n}^{-1}\mathbb{E}\|\boldsymbol{\mu}-\widehat{\boldsymbol{\mu}}\|^2 &= \widetilde{n}^{-1}\mathbb{E}\big\|\overline{\mathbf{W}}^{1/2}\mathbf{z}-\mathbf{P}_{\boldsymbol{\lambda}}\overline{\mathbf{W}}^{1/2}\mathbf{z}-\boldsymbol{\varepsilon}\big\|^2 \\ &= \widetilde{n}^{-1}\mathbb{E}\left[\big\|\overline{\mathbf{W}}^{1/2}(\mathbf{z}-\mathbf{D}\boldsymbol{\vartheta}^*)\big\|^2+\|\boldsymbol{\varepsilon}\|^2-2\big\langle\overline{\mathbf{W}}^{1/2}\mathbf{z}-\mathbf{P}_{\boldsymbol{\lambda}}\overline{\mathbf{W}}^{1/2}\mathbf{z};\boldsymbol{\varepsilon}\big\rangle\right] \\ &= \widetilde{n}^{-1}\mathbb{E}\big\|\overline{\mathbf{W}}^{1/2}(\mathbf{z}-\mathbf{D}\boldsymbol{\vartheta}^*)\big\|^2-1+2\widetilde{n}^{-1}\mathrm{tr}(\mathbf{P}_{\boldsymbol{\lambda}}), \qquad (2.11)\end{aligned}$$

where $\widetilde{n}:=5n$. The last line is recovered by expanding the inner product $\langle\cdot\rangle$, and by constructing $\overline{\mathbf{W}}^{-1/2}\mathbf{u}=:\boldsymbol{\varepsilon}\sim(\mathbf{0}_{\widetilde{n}},\mathbf{I}_{\widetilde{n}})$, the stochastic component of the GLS model leading to estimator (2.10). The trace of the influence matrix that appears in (2.11), computed by $\mathrm{tr}(\mathbf{P}_{\boldsymbol{\lambda}})=\mathrm{tr}(\mathcal{I}_{\mathrm{p}}^{-1}\mathcal{I})$, defines the effective degrees of freedom (edf) of the model. They usually differ from the number of parametric model components because of the presence of the penalty matrix which can suppress some dimensions of the parameter space. Multiple smoothing parameter selection can then be performed via minimisation of (2.11), an estimator that is commonly termed UBRE and that reads as

$$\begin{aligned}\boldsymbol{\lambda}^{[\alpha+1]} &:= \arg\min_{\boldsymbol{\lambda}}\mathcal{V}_u(\boldsymbol{\lambda}) \\ \mathcal{V}_u(\boldsymbol{\lambda}) &:= \big\|\overline{\mathbf{W}}^{1/2[\alpha+1]}(\mathbf{z}^{[\alpha+1]}-\mathbf{D}^{[\alpha+1]}\boldsymbol{\vartheta}^{[\alpha+1]}|_{\boldsymbol{\lambda}=\boldsymbol{\lambda}^{[\alpha]}})\big\|^2/\widetilde{n}-1+2\mathrm{tr}(\mathbf{P}_{\boldsymbol{\lambda}})|_{\boldsymbol{\lambda}=\boldsymbol{\lambda}^{[\alpha]}}/\widetilde{n}.\end{aligned}$$

Alternative ways to select $\boldsymbol{\lambda}$ can be defined starting from the working linear model (2.10): the GCV, for example, is also left as an option in our routine. The corresponding criterion is given explicitly by Wood (2006).

As previously anticipated, a link between (2.11) and the log-likelihood AIC exists. In fact, upon approximating $-2\ell(\boldsymbol{\vartheta}^*)$ about $\boldsymbol{\vartheta}$, it can be shown that

$$-2\ell(\boldsymbol{\vartheta}^*)\approx-2\ell(\boldsymbol{\vartheta})-\big\|\overline{\mathbf{W}}^{-1/2}\mathbf{u}\big\|^2+\big\|\overline{\mathbf{W}}^{1/2}(\mathbf{z}-\mathbf{D}\boldsymbol{\vartheta}^*)\big\|^2.$$

Hence, by realising that the smoothing parameter vector enters the above expression only through $\boldsymbol{\vartheta}^*$, dropping all irrelevant terms yields

$$\mathcal{V}_u(\boldsymbol{\lambda})\propto-2\ell(\boldsymbol{\vartheta}^*)+2\mathrm{tr}(\mathbf{P}_{\boldsymbol{\lambda}})=\mathrm{AIC}(\boldsymbol{\lambda}).$$

The steps described in these sections are made operative by adapting to the present context the outer iteration algorithm of O'Sullivan et al. (1986), which is detailed in Algorithm 1. In empirical analyses, however, the fitted tuning parameters may result in curves' estimates

Algorithm 1 Computation of the MPLE within a Trust-region Optimisation Routine

Require: $\alpha \in (0, \texttt{iter.max})$; $d \in [0, 1/4]$; $\bar{\Delta} > 0$; $\kappa \geq 1$
  $\boldsymbol{\vartheta}^{[0]}$, $\boldsymbol{\lambda}^{[0]}$, $\boldsymbol{p}^{[0]}$, $\Delta^{[0]} \in (0, \bar{\Delta})$
  while $\alpha \leq \texttt{iter.max}$ or $\max \left| \boldsymbol{\vartheta}^{[\alpha+1]} - \boldsymbol{\vartheta}^{[\alpha]} \right| \geq 10^{-6}$ **do**
    $\boldsymbol{p}^{[\alpha+1]} \leftarrow \min_{\boldsymbol{p}} - \left[\ell_{\mathrm{p}}(\boldsymbol{\vartheta}^{[\alpha]}|\boldsymbol{\lambda}^{[\alpha]}) + \boldsymbol{p}^\top \nabla_{\boldsymbol{\vartheta}^{[\alpha]}} \ell_{\mathrm{p}}(\boldsymbol{\vartheta}^{[\alpha]}|\boldsymbol{\lambda}^{[\alpha]}) + \frac{1}{2} \boldsymbol{p}^\top \nabla_{\boldsymbol{\vartheta}^{[\alpha]}\boldsymbol{\vartheta}^{[\alpha]\top}} \ell_{\mathrm{p}}(\boldsymbol{\vartheta}^{[\alpha]}|\boldsymbol{\lambda}^{[\alpha]}) \boldsymbol{p} \right]$
      s.t. $\|\boldsymbol{p}\| \leq \Delta^{[\alpha]}$
    $\varrho^{[\alpha+1]} \leftarrow \left[\ell_{\mathrm{p}}(\boldsymbol{\vartheta}^{[\alpha]}|\boldsymbol{\lambda}^{[\alpha]}) - \ell_{\mathrm{p}}(\boldsymbol{\vartheta}^{[\alpha]} + \boldsymbol{p}^{[\alpha+1]}|\boldsymbol{\lambda}^{[\alpha]}) \right] / \left[\tilde{\ell}_{\mathrm{p}}(\mathbf{0}|\boldsymbol{\lambda}^{[\alpha]}) - \tilde{\ell}_{\mathrm{p}}(\boldsymbol{p}^{[\alpha+1]}|\boldsymbol{\lambda}^{[\alpha]}) \right]$
    if $\varrho^{[\alpha+1]} < 1/4$ **then**
      $\Delta^{[\alpha+1]} \leftarrow 1/4 \Delta^{[\alpha]}$
    else
      if $\varrho^{[\alpha+1]} > 3/4$ and $\|\boldsymbol{p}^{[\alpha+1]}\| = \Delta^{[\alpha]}$ **then**
        $\Delta^{[\alpha+1]} \leftarrow \min(2\Delta^{[\alpha]}, \bar{\Delta})$
      else
        $\Delta^{[\alpha+1]} \leftarrow \Delta^{[\alpha]}$
    if $\varrho^{[\alpha+1]} > d$ **then**
      $\boldsymbol{\vartheta}^{[\alpha+1]} \leftarrow \boldsymbol{\vartheta}^{[\alpha]} + \boldsymbol{p}^{[\alpha+1]}$
    else
      $\boldsymbol{\vartheta}^{[\alpha+1]} \leftarrow \boldsymbol{\vartheta}^{[\alpha]}$
    $\boldsymbol{\lambda}^{[\alpha+1]} \leftarrow \min_{\boldsymbol{\lambda}} \left[\left\| \overline{\mathbf{W}}^{1/2[\alpha+1]} (\mathbf{z}^{[\alpha+1]} - \mathbf{D}^{[\alpha+1]} \boldsymbol{\vartheta}^{[\alpha+1]}|_{\boldsymbol{\lambda}=\boldsymbol{\lambda}^{[\alpha]}}) \right\|^2 / \widetilde{n} - 1 + 2\kappa \mathrm{tr}(\mathbf{P}_{\boldsymbol{\lambda}}|_{\boldsymbol{\lambda}=\boldsymbol{\lambda}^{[\alpha]}}) / \widetilde{n} \right]$

that are believed to be too wiggly by the researcher. If that is the case, the trace of the influence matrix can be inflated by a scaling parameter $\kappa \geq 1$ to obtain smoother functions. We refer the reader to Kim and Gu (2004) for more details on this point.

2.3.4 Further Results and Inference

At convergence of the optimisation algorithm, point-wise confidence intervals for the estimated non-parametric curve $\widehat{s}_{j,l_j}$ can be obtained from the distribution

$$\mathcal{N}(s_{j,l_j}(\mathrm{v}_{j,l_j,i}), \boldsymbol{b}_{j,l_j,i}^\top \mathbf{V}_{\boldsymbol{\vartheta},[j,l_j]} \boldsymbol{b}_{j,l_j,i}),$$

where $\mathbf{V}_{\boldsymbol{\vartheta},[j,l_j]}$ denotes the sub-matrix of $\mathbf{V}_{\boldsymbol{\vartheta}}$ corresponding to the parameters associated to the (j,l_j)-th smooth, and $\mathbf{V}_{\boldsymbol{\vartheta}} := -\mathcal{H}_{\mathrm{p}}^{-1}$ is the covariance matrix of the posterior distribution of $\boldsymbol{\vartheta}|\mathbf{w} \sim \mathcal{N}_p(\widehat{\boldsymbol{\vartheta}}, \mathbf{V}_{\boldsymbol{\vartheta}})$, with $\mathbf{w} := \mathbf{D}^\top \overline{\mathbf{W}} \mathbf{z}$. For the smooth functions included in the model $\mathbf{V}_{\boldsymbol{\vartheta}}$ is usually preferred to the more intuitive estimator $\mathbf{V}_{\widehat{\boldsymbol{\vartheta}}} := -\mathcal{H}_{\mathrm{p}}^{-1} \mathcal{H} \mathcal{H}_{\mathrm{p}}^{-1}$. In fact, as Marra and Wood (2012) showed in the context of GAMs, the former includes both a bias and a variance components in a frequentist sense, a feature that is not shared by $\mathbf{V}_{\widehat{\boldsymbol{\vartheta}}}$.

The construction of the posterior distribution above was firstly advocated by Wahba (1983) and Silverman (1985). They recognised that any penalised estimation framework has

a natural counterpart in the explication of some prior believes about the likely features of the true model. In particular, the imposition of a conjugate normal prior for $\boldsymbol{\vartheta}$ assumes that smoother models are more probable than wiggly ones, whilst the same probability density is assigned to all models of equal smoothness. Therefore, combining this reasoning with the normality of $\mathbf{w}$ (Wood, 2006), the stated result emerges.

Confidence intervals for non-linear functions of the parameter vector $\boldsymbol{\vartheta}$ can also be constructed using a convenient simulation scheme from the posterior distribution of $\boldsymbol{\vartheta}|\mathbf{w}$. We articulate the corresponding algorithm as follows. Let $T(\boldsymbol{\vartheta})$ be any function of the parameters, then

1. draw N_{sim} vectors $\boldsymbol{\vartheta}_r^*, r = 1, \ldots, N_{sim}$, from $\mathcal{N}_p(\widehat{\boldsymbol{\vartheta}}, \mathbf{V}_{\boldsymbol{\vartheta}}(\widehat{\boldsymbol{\vartheta}}))$, where $\widehat{\boldsymbol{\vartheta}}$ is the MPLE;
2. compute $T_r^* := T(\boldsymbol{\vartheta}_r^*)$ for every r, and define T_α^* to be the $[N_{sim}\alpha]$-th smallest value of the ordered sample $\{T_{(1)}^*, \ldots, T_{(N_{sim})}^*\}$, with $[a]$ denoting the integer part of $a \in \mathbb{R}$;
3. obtain an approximate $(1-\alpha)\%$ confidence interval for $T(\widehat{\boldsymbol{\vartheta}})$ using $[T_{\alpha/2}^*, T_{1-\alpha/2}^*]$.

To gain insights into the effectiveness of the estimation approach, the results from a small Monte Carlo simulation study are presented in Figure 2.1. For the sake of conciseness, the exact definition of the Data Generating Process (DGP) is provided in Appendix B.1. On average, the experiment shows that our method appears to be effective in recovering the true functions, although with a higher degree of uncertainty for the smooth in the simulated equation of Y_2 (last panel in the figure). This result is not unexpected. The recursive formulation of the model implies that the curves defining Y_1 enter the second equation directly through reduced-form system (2.2). Hence estimation of the corresponding parameters has to account also for this further source of uncertainty which stems from the first equation of the model. The same experiment has been repeated for $n = 3,000$ after the suggestion of one reviewer (see Appendix B.1). A similar pattern of Figure 2.1 is maintained when the sample size is reduced, although the uncertainty in recovering the curves is more evident in this case.

Some Asymptotic Considerations The large sample behaviour of the MPLE can be established under the relatively mild conditions of the consistency of the MLE. Following the arguments of Kauermann (2005), let us define

$$\boldsymbol{\vartheta}_0 := \arg\min_{\boldsymbol{\vartheta}} \mathrm{KL}(\mathcal{L}_t|\mathcal{L}_n) = \mathbb{E}[\ell_t - \ell_n(\boldsymbol{\vartheta})]$$

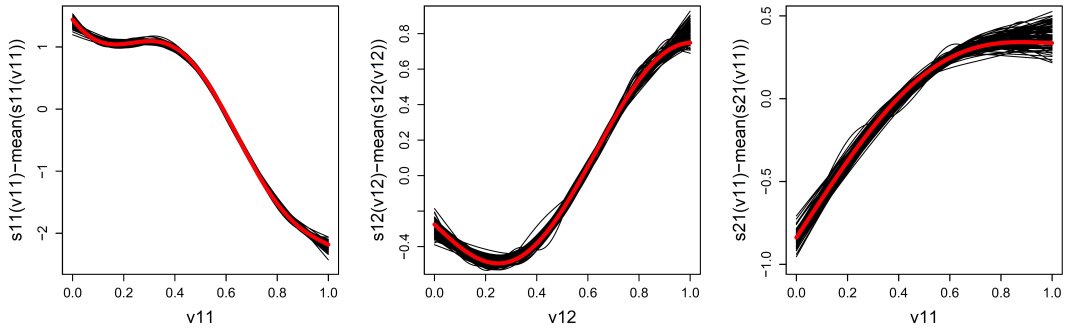

Figure 2.1: Estimated smooth curves obtained from 100 replicates of a Monte Carlo experiment comprising 10,000 simulated observations (true curves in red). Parameters' values were set close to the ones recovered in fitting the empirical illustration, in particular we have defined $\psi = -0.3$ and $\rho = 0.2$. The smooth components were represented using penalized thin plate regression splines with basis dimensions equal to 10 and penalties based on second-order derivatives. Results are plotted on the scale of the linear predictors. Please refer to Appendix B.1 for the exact definition of the DGP employed.

be the minimiser of the Kullback-Leibler discrepancy between the true structure that has generated the data and the employed model, and set the spline bases at a fixed high dimension. This is a rather convenient assumption, but still of some relevance in applied research where the bases' dimension has to be fixed in order to achieve estimation. An existing drawback, however, is that the unknown smooth functions may not have an exact representation as linear combinations of the given bases at a finite dimension. Hence they may not be asymptotically recovered by their estimators as the sample size increases. Nonetheless, by using a number of bases rich enough to obtain a good representation of the unknown curves, it is possible to assume heuristically that the approximation bias is negligible compared to estimation variability (Kauermann, 2005).

Further let the following conditions hold: (i) $\nabla_{\boldsymbol{\vartheta}_0}\ell_n = O_p(n^{1/2})$, (ii) $\mathbb{E}[\nabla_{\boldsymbol{\vartheta}_0\boldsymbol{\vartheta}_0^\top}\ell_n] = O(n)$, (iii) $\nabla_{\boldsymbol{\vartheta}_0\boldsymbol{\vartheta}_0^\top}\ell_n - \mathbb{E}[\nabla_{\boldsymbol{\vartheta}_0\boldsymbol{\vartheta}_0^\top}\ell_n] = O_p(n^{1/2})$, and (iv) $\mathbf{S}_{\boldsymbol{\lambda}} = o(n^{1/2})$. Assumptions (i)-(iii) are the usual ones for the MLE consistency, whereas the last one is equivalent to consider $\lambda_{j,m_j} = o(n^{1/2})$ for any j. This comes from the very construction of the penalty matrix, and from the fact that every $\mathbf{S}_{j,m_j}$ is asymptotically bounded. Then the MPLE can be proved to satisfy

$$\widehat{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta}_0 = \mathbf{F}^{-1}(\boldsymbol{\lambda})(\nabla_{\boldsymbol{\vartheta}_0}\ell(\boldsymbol{\vartheta}_0) - \mathbf{S}_{\boldsymbol{\lambda}}\boldsymbol{\vartheta}_0)[1 + o_p(1)], \tag{2.12}$$

where $\mathbf{F}^{-1}(\boldsymbol{\lambda}) = (\mathbf{S}_{\boldsymbol{\lambda}} - \mathbb{E}[\nabla_{\boldsymbol{\vartheta}_0\boldsymbol{\vartheta}_0^\top}\ell(\boldsymbol{\vartheta}_0)])^{-1}$, and the leading stochastic component in (2.12) has asymptotic order $O_p(n^{-1/2})$ as $n \to \infty$. The proof of this result is given in Appendix A.1.

2.4 The Effect of Education on Drinking Behaviour in the UK

Alcohol misuse has serious effects on global health and is commonly regarded as the third major risk factor for premature deaths and disabilities in the world (World Health Organization, 2007). It is also linked to a number of pathological conditions (e.g., coronary heart disease, stroke, liver disease and various cancers). The level of alcohol consumption in the United Kingdom has been recently reported by the OECD to be above the average of the EU countries (10.6 liters per capita against an average of 10.1 in 2012) and, despite its gradual decline in the EU from 1980, it has remained stable in the UK since then (OECD, 2014). In a report by the Public Health England, the HM Government acknowledged that as many as 21,485 people died in 2012 from alcohol-related causes out of a total of around nine-million adults who drink at levels that pose some risk to their health (Public Health England, 2014). This comes with high costs for the society too. It has been estimated a total annual cost of alcohol-related harm of 21bn GBP, with an impact of 3.5bn GBP a year in costs related to alcohol for the National Health Service (NHS). The harmful use of alcohol compromises both individual and social development. The Crime Survey for England and Wales 2012-13, for example, showed that 49% of all violent crimes was connected to alcohol, with peaks involving 69% of stranger and 38% of domestic violences. In addition, problem drinking by parents is thought to contribute to the development of physical, psychological and behavioural problems in children.

In this study, we aim at applying the ideas discussed in the chapter to investigate the effect of education on alcohol consumption in Great Britain. This is a non-trivial relationship since the level of education can act at different levels, and its overall effect is theoretically ambiguous. Recently, Huerta and Borgonovi (2010) surveyed and further elaborated on this aspect. On the one hand, more educated individuals are argued to have access to a wider spectrum of information relating to healthy behaviours, and usually acquire the necessary skills to process them and to act accordingly (Brunello et al., 2008, Goldman and Smith, 2005). Hence they may have a deeper knowledge about the risks connected to alcohol abuse (Kenkel, 1991). On the other hand, however, education shapes labour market opportunities and the social context in which people operate. As a result, better educated individuals face in general fewer financial constraints and may be exposed to working environments where drinking is acceptable if not even expected. Alongside with this lack of social stigma, an active social life and a high sense of self-control may lead these people to have more frequent and possibly heavier drinking sessions than those of their less educated peers.

In addition to these conflicting directions in the sought relation, a number of other studies have also acknowledged the relevance of the time preference in predicting alcohol consumption (see O'Donoghue and Rabin, 2000, Fehr, 2002 and Delaney et al., 2008 just to name but a few). In particular, they indicate that people generally show a high rate of time preference with respect to their drinking behaviour, so that it is commonly perceived to be myopic. In other words, individuals tend to be more willing to put the well-being deriving from alcohol intake in the present rather than in the future, and this occurs at the expense of possible health-related problems. Education is also well understood to be associated with time preference. This point has been raised by Sander (1995), Bratti and Miranda (2010) and references therein in the context of smoking margins, and by Fuchs (1982) and van der Pol (2011) in a more general setting. Disentangling the true association between education and drinking behaviour requires therefore to account for this possible source of omitted variable bias. In the words of van der Pol (2011): "both education decisions and health decisions involve trade-offs of outcomes over time. Individuals' time preferences [...] will therefore influence how individuals make intertemporal choices such as whether or not to invest in education, whether to save or borrow and whether to engage in health affecting behaviours such as smoking, drinking and drug use" (p. 917). Finally, combining the evidence of low time preference for the *choice of education* and the aforementioned myopic *attitude towards alcohol consumption*, one would expect individuals' time preference to drive the latent counterparts of the two variables of interests in opposite directions which corresponds, in the current formulation of the model, to a negative correlation coefficient, $\rho < 0$.

2.4.1 Data and Empirical Analysis

We fit the simultaneous equation system model proposed in this chapter to data from the 1970 British Cohort Study (BCS70), a longitudinal dataset of all children born in the Great Britain from the 5th to the 11th of April 1970, for a total of 17,198 babies surveyed. Information on the maximum educational level attained by the participants, as well as data on their geographical location and drinking behaviour were collected in the 29-year follow-up survey, whereas all the remaining variables are from the 10-year follow-up. This choice has been made primarily for data availability and the lower level of attrition experienced at these waves: after a first screening of the answers, we have a sample size of 7,115 respondents against the original 10,405 as from the merging of the two waves considered. Notice that item non-response in our main drinking variable, self-reported quantity of alcohol intake in

Highest Education	Alcohol Consumption					Marginals
	1	2	3	4	5	
Up to O-levels	1,191 (16.74%)	733 (10.30%)	1,125 (15.81%)	401 (5.64%)	1,285 (18.06%)	4,735 (66.55%)
A-levels	91 (1.28%)	71 (1.00%)	123 (1.73%)	47 (0.66%)	137 (1.93%)	469 (6.59%)
Higher Education	286 (4.02%)	184 (2.59%)	553 (7.77%)	218 (3.06%)	670 (9.42%)	1,911 (26.86%)
Marginals	1,568 (22.04%)	988 (13.89%)	1,801 (25.31%)	666 (9.36%)	2,092 (29.40%)	7,115 (100.00%)

Table 2.2: Empirical distribution of the observed categories for the response variables in the BCS70 29-year follow-up. In brackets we have reported the corresponding fraction of the sample size. Alcohol consumption is categorical in the original survey and is represented here with levels ranging from 1: "less often/only on special occasions (1,414); never nowadays (399); never had an alcoholic drink (192); don't know (4); not answered (16)" to 5: whoever drinks above the NHS recommended limits. Notice that level 1 includes also those individuals who declared themselves to drink at least once in a week, but no information about amount of alcohol consumed is reported (322).

the week prior to the interview, is very low (30), whereas a higher proportion of incomplete responses (2,090) were collected for the British Ability Scales (BAS). This is a battery of cognitive and achievement tests submitted to individuals and accounted in the 10-year follow-up.

The corresponding empirical bivariate densities of the dependent variables of interest, "highest education achieved" and "alcohol consumption", are given in Table 2.2. We note that the majority of respondents attended at most the O-levels, the compulsory lower secondary educational qualification in the UK, whilst only few people completed the A-levels without proceeding to any kind of Higher Education (HE). Concerning alcohol intake, around 39% of cohort members had alcoholic drinks in a week time at a level (in terms of units) of potential harm for their health. This threshold has been set according to the NHS recommendations of 2-3 units a day for women, and 3-4 for men. After having translated the different types of beverages into the corresponding alcohol units, we have distinguished usual drinkers between whoever intakes units within the suggested weekly limits ($\leq$ 14 u/w, level 3), "just" in the limits (14-21 u/w, level 4), and above them ($>$ 21 u/w, level 5). The values provided refer to women and the corresponding amounts for men can be computed analogously from the daily NHS recommendations. The remaining levels 1 and 2 comprise people who declared themselves to be only occasional/not drinkers at all, and light drinkers, respectively.

Our model specification follows the one proposed in the literature by Bratti and Miranda (2009) and Huerta and Borgonovi (2010) in a similar context, and controls for some childhood circumstances that are commonly associated with alcohol abuse (Caldwell et al., 2008,

Droomers et al., 2003, Hemmingsson et al., 1999, Poulton et al., 2002). In particular, we include variables referring to the parental presence in children's life and their interest in children's education, maternal weekly working hours, the highest parental social class, ethnicity and home tenure. The precise definition of these variables, along with their corresponding labels in the dataset, are given in Appendix B.1 for replication. We have excluded from the equation of the alcohol consumption the score obtained by the respondents in the BAS at the age of 10. In fact, this variable is generally understood to affect the highest level of education attained by the cohort members; nonetheless, it is also unlikely that the results of a test sat at an early age can be a *direct* predictor of the quantity of alcohol intake or drinking frequency at the age of 29, but through its effects on educational achievements. The same variable was also excluded by Bratti and Miranda (2009, 2010) in studying the effect of education on drinking frequency and smoking intensity in a similar bivariate framework. The system of equations, in `R` notation, is then:

$$\begin{aligned} \texttt{edu}_i^* \;\sim\; & \texttt{mum.not.pres}_i + \texttt{dad.not.pres}_i + \texttt{mum.edu}_i + \texttt{dad.edu}_i + \texttt{s.class}_i + \\ & \texttt{eth.child}_i + \texttt{mum.int.edu}_i + \texttt{dad.int.edu}_i + \texttt{sex.b}_i + \texttt{home}_i + \\ & s(\texttt{mum.wrk.hr}_i) + s(\texttt{BAS.tot}_i) \end{aligned}$$

$$\begin{aligned} \texttt{drk5}_i^* \;\sim\; & \texttt{edu}_i^* + \texttt{mum.not.pres}_i + \texttt{dad.not.pres}_i + \texttt{mum.edu}_i + \texttt{dad.edu}_i + \texttt{s.class}_i + \\ & \texttt{eth.child}_i + \texttt{mum.int.edu}_i + \texttt{dad.int.edu}_i + \texttt{sex.b}_i + \texttt{home}_i + \texttt{region}_i + \\ & s(\texttt{mum.wrk.hr}_i), \end{aligned}$$

where the categorical covariates are included in the above formulae as `as.factors(·)`.

Results and Interpretation The estimated parameters obtained by employing the above model are given in Table 2.3 for any discrete covariate, and in Figure 2.2 for the continuous predictors. For comparison purposes, the same specification has also been used for the fully parametric version of the proposed model, whose estimates are only reported for ψ and ρ as the main parameters of interest.

Although the raw estimates are not interpretable *per se*, a quick assessment of the converged log-likelihoods shows that some gains are indeed achieved by employing a semi-parametric model rather than just assuming linear covariate effects. At the same time, we note that the fitted values obtained from the two models are very close to each other. A possible explanation is that the effects of maternal weekly working hours are either esti-

TRIANGULAR SEMI-PARAMETRIC BIVARIATE ORDERED PROBIT REGRESSION														
Variables	Highest Education Achieved							Alcohol Consumption						
	Covariates' levels													
	level 2	level 3	level 4	level 5	level 6	level 7	level 8	level 2	level 3	level 4	level 5	level 6	level 7	level 8
Mother not present	-.605	.112	–	–	–	–	–	.125	-.111	–	–	–	–	–
	(1.031)	(.098)	–	–	–	–	–	(.844)	(.078)	–	–	–	–	–
Father not present	.747	.093	–	–	–	–	–	-.187	.062	–	–	–	–	–
	(.819)	(.055)	–	–	–	–	–	(.678)	(.045)	–	–	–	–	–
Mother highest education	-.101	.175	.238	.482	.514	–	–	-.065	.025	-.015	-.004	.148	–	–
	(.069)	(.042)	(.081)	(.072)	(.107)	–	–	(.054)	(.037)	(.072)	(.066)	(.095)	–	–
Father highest education	.119	.127	.229	.151	.492	–	–	.102	.086	.021	.047	-.085	–	–
	(.066)	(.047)	(.057)	(.096)	(.058)	–	–	(.055)	(.039)	(.050)	(.085)	(.055)	–	–
Social class	-.112	-.125	-.365	-.108	-.314	-.581	–	-.042	.021	-.085	-.054	-.053	-.023	–
	(.118)	(.057)	(.075)	(.072)	(.058)	(.128)	–	(.095)	(.049)	(.062)	(.062)	(.050)	(.094)	–
Ethnicity	-.688	-.783	–	–	–	–	–	.523	.682	–	–	–	–	–
	(.059)	(.098)	–	–	–	–	–	(.473)	(.090)	–	–	–	–	–
Mother interested in child's education	-.050	-.029	-.212	–	–	–	–	.027	.008	-.109	–	–	–	–
	(.098)	(.059)	(.202)	–	–	–	–	(.078)	(.048)	(.127)	–	–	–	–
Father interested in child's education	-.101	-.113	-.180	–	–	–	–	.060	.050	.030	–	–	–	–
	(.063)	(.040)	(.166)	–	–	–	–	(.050)	(.032)	(.112)	–	–	–	–
Gender	-.089	–	–	–	–	–	–	.647	–	–	–	–	–	–
	(.032)	–	–	–	–	–	–	(.026)	–	–	–	–	–	–
Home tenure	-.028	-.406	-.102	–	–	–	–	.072	.160	.168	–	–	–	–
	(.190)	(.050)	(.044)	–	–	–	–	(.164)	(.044)	(.038)	–	–	–	–
Region of residence (levels 9-11 omitted)	–	–	–	–	–	–	–	-.079	.006	.021	-.459	-.080	.079	.032
	–	–	–	–	–	–	–	(.046)	(.054)	(.051)	(.290)	(.069)	(.043)	(.053)
	semi-parametric												parametric	
ψ	0.300												0.304	
	(0.023)												(0.022)	
ρ	-0.217												-0.221	
	(0.032)												(0.031)	
log-likelihood	-15,192.81												-15,207.73	
No. observations	7,115												7,115	

Table 2.3: Estimated parameters for the categorical covariates included in the proposed triangular semi-parametric probit model; standard errors are reported in round brackets under the corresponding estimates. A comparison between the regression splines and the purely parametric models is included at the bottom.

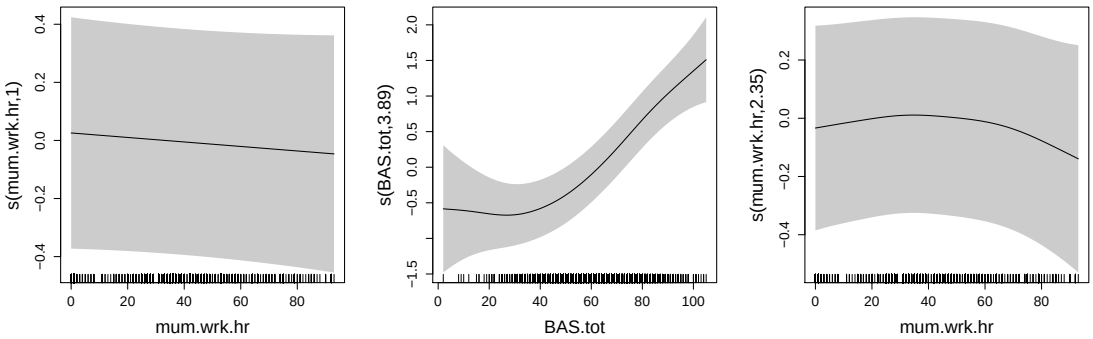

Figure 2.2: Estimated smooth functions and associated 95% point-wise confidence interval obtained by applying `SemiParCLM` to the BCS70 dataset. The first two curves correspond to the functions included in the equation for the educational achievements, while the last one to the model for the drinking frequency. The effective degrees of freedom are reported into brackets in the y-axis caption, with a value of one denoting the estimation of a straight line (as for the first curve). The actual covariate values are reported at the bottom of each graph through a jittered rug plot. The functions have been estimated using a low-rank penalized thin plate regression spline with basis dimensions equal to 10 and penalties based on second-order derivatives.

mated as a straight line, or are not important predictors for the responses. This conclusion is drawn from the observation that the zero line is entirely contained within the confidence intervals of the smooths. Hence, the mis-specification bias induced by a parametric functional form seems to be less amplified in this particular application. Quite interestingly, after having controlled for the possible source of omitted variables in the study, we find that childhood circumstances do not tend to be explanatory of the determination of both educational achievements and alcohol consumption of the cohort members. On the other hand, usual socio-demographic characteristics like parental education, social class, ethnicity and home tenure contribute to the explanation of children's highest level of education. This pattern is also confirmed by the estimated non-parametric curves, which appear to be uninformative in predicting the corresponding responses apart from the BAS values. To further check on this, a shrinkage approach to variable selection in the spirit of Marra and Wood (2011) was performed (results are reported in Appendix B.1 for the sake of space). This method highlights that maternal working hours is not an influential predictor for the first equation, and hence it is safe to drop it from the current model specification. Final results remain, however, unchanged.

As previously anticipated, we can actually comment on the finding of a negative correlation among the two latent variables of the bivariate model. Specifically, if time preference is assumed to drive together the choice of education and the consumption of alcoholic bev-

erages, the latter through its effect on undertaking healthier behaviours, then people who decide to invest in more schooling are also those less incline at recognising (or considering) the consequences of alcohol abuse on their future health status. The estimated correlation coefficient is statistically different from zero, with a reported p-value for the null hypothesis $H_0 : \rho = 0$ of < 0.000. Therefore, the use of a simple univariate model which does not correct for the possible presence of omitted variables in the association of interest would have resulted in inconsistent estimates.

To give a better picture of the situation, we investigate the effects of education on people's weekly units of alcohol intake by looking at the predicted conditional probabilities (e.g. Greene and Hensher, 2010). Namely, we compute the probability of the average individual to consume a certain quantity of alcohol given his/her observed educational achievements. Formally, we have

$$\widehat{\mathrm{PP}}_{k_2|k_1,i} := \frac{\mathbb{P}[y_{1,i} = k_1, y_{2,i} = k_2]}{\mathbb{P}[y_{1,i} = k_1]} = \left. \frac{\sum_{l,m\in\{0,1\}}(-1)^{l+m}\Phi_2(\eta_{1,k_1,i}(\boldsymbol{\vartheta}), \eta_{2,k_2,i}(\boldsymbol{\vartheta}); \boldsymbol{\Sigma})}{\Phi(\eta_{1,k_1,i}(\boldsymbol{\vartheta})) - \Phi(\eta_{1,k_1-1,i}(\boldsymbol{\vartheta}))} \right|_{\boldsymbol{\vartheta}=\widehat{\boldsymbol{\vartheta}}},$$

and the corresponding average effect is consequently $\widehat{\mathrm{APP}}_{k_2|k_1} = n^{-1}\sum_i \widehat{\mathrm{PP}}_{k_2|k_1,i}(\widehat{\boldsymbol{\vartheta}})$. The confidence intervals can be computed using the simulation approach detailed in Section 2.3.4. Table 2.4 finally reports the $\widehat{\mathrm{APP}}$ values for every combination of k_1 and k_2. In line with the theoretical arguments provided in the literature, we find that individuals with a HE qualification have a larger probability to intake weekly alcohol units above the NHS recommendations, and to drink more often than the lesser educated ones. The latter has been established by replacing the main drinking variable with a new one measuring the frequency of alcohol consumption during the week. Results are given in Appendix B.1. In particular, the "degree effect" accounts for a 5.28% higher probability to drink at harmful levels compared to individuals who have (at most) completed the compulsory schooling. However, less education tends to be associated with a higher probability of being an occasional and/or a light alcohol consumer by 5.65% (2.05% if the individual got A-levels) with respect to people with a university degree.

As a note of caution, we warrant that the results obtained may not lend to an immediate generalisation to other contexts due to the very nature of the data analysed. In fact, although alcohol consumption is often regarded to vary with location and age, among other factors, cohort members were all born in the same week of 1970 in the UK, and their relevant drinking variables referring to the 29-year follow-up. Nonetheless, the reported association reveals,

Highest Education	Alcohol Consumption				
			at least one drink per week		
	no/occasional	light	$<$ NHS limits	$\approx$ NHS limits	$>$ NHS limits
Up to O-levels	0.2320	0.1424	0.2550	0.0926	0.2780
	(.1523; .3304)	(.0432; .2230)	(.2454; .2648)	(.0833; .1016)	(.2698; .2858)
A-levels	.2029	0.1354	0.2554	0.0967	0.3095
	(.1300;.2957)	(.0417; .2093)	(.2457; .2651)	(.0870; .1061)	(.3010; .3181)
HE or equivalent	0.1876	0.1303	0.2530	0.0983	0.3307
	(.1182;.2771)	(.0405; .1998)	(.2434; .2627)	(.0885; .1079)	(.3243; .3375)

Table 2.4: Average predicted conditional probabilities: each entry indicates the probability of a randomly drawn individual to have a certain weekly quantity of alcohol intake given his/her observed highest educational achievement. The 95% confidence intervals reported below the estimates are computed through simulation from the posterior distribution of $\boldsymbol{\vartheta}|\mathbf{w}$.

from a policy standpoint, that a raise in alcohol duties may not affect its (mis)consumption by the social group of educated young adults. In fact, because of the significant monetary wage returns of Higher Education (as documented, for example, by Blundell et al. 2000), this group may tend to be less price elastic, with a demand which is less responsive to a price change.

2.5 Concluding Remarks

In this chapter we have introduced a bivariate triangular ordinal probit regression with semi-parametric covariate effect. Our model formulation has been recovered as an instance of a penalized Generalized Linear Model framework, so that estimation and inference have been conducted as a natural extension of GLMs. Semi-parametric modelling is of relevance in applications as it allows the researchers to achieve a higher degree of flexibility in empirical modelling. Hence it alleviates the bias arising from model mis-specification.

Following some relevant examples given in the literature (e.g. Sajaia 2008 and Buscha and Conte 2014), we have defined a prototypical recursive model for ordinal polychotomous responses collected in $\boldsymbol{Y} = (Y_1, Y_2)^\top$. This specification is usually employed in observational studies to account for the possible presence of unobserved confounding. Specifically, we have assumed that a response Y_2 of interest (alcohol consumption in the empirical illustration) is structurally dependent on a variable Y_1 (education achievements), and that a third factor affecting simultaneously the two is omitted from the analysis because not readily quantifiable (e.g., individual time preferences). In general, such an omission may induce a further source of association between Y_1 and Y_2 which is different from the relationship that the researcher is willing to investigate. This fact has been accounted for by estimating a correlation parameter

that captures the association implied by the confounder(s). Furthermore, we have identified the relationship between the elements in $\boldsymbol{Y}$ by including a variable which is independent of the time preference, does not affect the intake of alcohol units at the age of 29 (holding the educational achievements constant) and that is relevant in predicting the highest education of cohort members. These conditions define what is commonly regarded as an exclusion restriction in econometrics and epidemiology.

Incidentally, we have also illustrated how the triangular representation can be further qualified to recover other models nested in it, as well as the required modifications to be made in case one of the Y_j's is dichotomous. However, some directions remain to be explored. In particular, it could be of interest to investigate to what extent the representation proposed is applicable to different mixtures of discrete responses' types beyond the dichotomous/ordinal polychotomous one, or to extend the system of equations to encompass more than two dimensions. The further specification of the correlation coefficient as a function of some covariates is also of interest. This might help to investigate the role of unmeasured confounders in more depth, and to relate them to specific variables. The possible extension of the approach of Gertheiss and Tutz (2009) to the present context would also be useful to incorporate the implied monotonicity of the ordered covariates in our estimation algorithm. We will address these issues in future research.

Chapter 3

Copula-based Approach to Penalized Likelihood Estimation of Car Accident Injuries

A bivariate system of equations is developed to model ordinal polychotomous dependent variables within an additive regression framework. The functional form of covariate effects is assumed fairly flexible, with appropriate smoothers included to account for non-linearities and spatial variability in the data. Non-Gaussian error dependence structures are dealt with using Archimedian copulae. The framework is then employed to study the effects of several risk factors on the levels of injury sustained by individuals in vehicle accidents in France. The use of a bivariate model is motivated by the possible presence of common unobservables that may affect the inter-relationships between the various parties involved in the same crash. In this way, more efficient estimates are obtained and mis-specification reduced via an enhanced model specification.

3.1 Introduction

Vehicle-related injuries are a source of major concern for national governments and international organizations as they impact the life of millions of individuals around the world. In a recent report, the World Health Organization estimated that around 1.24 million of people die in road accidents every year, whereas approximately 20/50 million are involved in non-fatal injuries (World Health Organization, 2013). This makes car crashes the eighth leading cause of death, and the prominent one for young people aged 15-29 years. A global awareness campaign on this issue has been launched by the United Nations General Assembly with resolution 64/255 which proclaimed the Decade of Action for Road Safety for the period 2011-2020. Its goal is to stabilise, and possibly reduce, the trend in road traffic fatalities and thus to save around 5 million lives over the foreseen action time.

To tackle this preventable major source of injury, several developed countries have created ad-hoc agencies funded from their national budgets. In France, for instance, a national task

force (Comité Interministériel à la Sécurité Routière) was established in 1972 with the aim of defining governmental policies on the matters of road safety and ensuring their proper and timely enforcement. Legislations and information are commonly recognised to play a significant role in prevention and, at least in high-income countries, they showed a generalised reduction in fatal injuries. Despite these encouraging results, however, the annual costs of crash injuries for society are still high and have been estimated to exceed 180 billion EUR in the European Union alone. Deaths too constitute a non-negligible figure in national statistics (World Health Organization, 2004). The French Observatoire National Interministériel de la Sécurité Routière (ONISR) counted in its "Baromètre du mois de juillet 2015" that as much as 3,384 people died within 30 days from any road injury in the country during the past year. A deeper understanding of crash causations and injury severities is therefore fundamental to improve roadway safety, and hence contribute to a transport system that is sustainable in terms of its economic and social costs.

The study of injury severities in vehicle crashes may nonetheless present several difficulties because of the intrinsic complexity of the problem. In fact, severity levels are often the result of many observed factors (e.g., road geometry, vehicle standards, behaviour of road users) and some others that can be hardly measured by data collectors. For example, speed before impact, presence of moving obstacles on the pavement, or even sudden environmental-related factors are not typically recorded properly by police officers at the time they arrive, whilst these may constitute important variables in accident dynamics. Acknowledging these concerns, in this chapter we develop a class of models accounting for the role that unobserved factors may play in the determination of injury severities in vehicle crashes. Specifically, this work discusses a bivariate Cumulative Link Additive Model as a semi-parametric extension of the family of the Cumulative Link Models (CLMs) originally formalised by McCullagh (1980). Notably, the effects of continuous covariates on the responses of interest are estimated using penalized regression splines. Hence non-linearities can be handled flexibly, without introducing, for instance, arbitrary categorisations of the relevant regressors into groups based on intervals or frequencies. In line with some recent methodological advances (e.g. Radice et al., 2015), we extend the prototype bivariate ordered probit regression of Chapter 2 to incorporate several dependence structures of the responses as induced by the class of Archimedean copulae. This constitutes an advantage in empirical studies: copulae allow us to specify models beyond the Gaussian distribution and employ different marginals irrespective of the particular association linking them. Importantly, researchers will be given

new tools to assess the sensitivity of their results under different model specifications and assumptions.

Copulae have been previously considered in the general transportation literature by Bhat and Eluru (2009), and by Eluru et al. (2010) and Rana et al. (2010) in the context of road safety. However, their treatment in the setting of ordinal polychotomous responses with non-parametric covariate effects has not been analysed yet. This paper aims therefore at filling this gap. Our estimation approach takes advantage of the multivariate penalized Generalized Linear Model (GLM, Nelder and Wedderburn, 1972) representation to be discussed in Chapter 4, in which various penalisation terms are used to enforce certain desired characteristics of the functional form of the covariate effects. Among them, non-linearities and spatial variation within the data have been shown to be all representable within that generic framework through the use of smoothers appropriately defined. These features are made operative here and automatically estimable by the function `CopulaCLM`, which implements the ideas discussed in this chapter for the `R` computational environment (R Development Core Team, 2015). In this respect, we extend to ordinal response levels the work on bivariate copula modelling with dichotomous outcomes of Radice et al. (2015). In the Bayesian literature, an analogous model for ordinal dependent variables has been introduced by Hillmann et al. (2014) for a neuroscience application. Their paper, however, only investigates the effects of Gaussian distributions on the classification ability of seemingly unrelated regression (SUR) equations, without assessing the impacts of different dependence structures.

Semi-parametric models have received scarce attention in accident research, despite their known superiority over traditional linear regressions in terms of model specification as well as estimates' efficiency. For example, it seems that the class of Generalized Additive Models (GAMs, Hastie and Tibshirani, 1986, 1990) has been only applied twice (Xie and Zhang, 2008 and Li et al., 2011). We suspect that this practice may be due to the predominant empirical interest in ordinal response models, for which some extra considerations are required to harmonise their structures with those of univariate GAMs (e.g. Yee and Wild, 1996). In this vein, our contribution is to introduce flexible tools which researchers can use to better assess the effects of observed covariates on the responses of interest, and allow for a more cautious judgment of the results obtained. In the words of Mannering and Bhat (2014), the "use of methodological approaches with known deficiencies [...] has the potential to lead to erroneous and ineffective safety policies that may result in unnecessary injuries and loss of life" (p. 16).

The remainder of the chapter is structured as follows. In Section 3.2 we introduce the statistical model and discuss its representation and main features. An empirical-oriented motivation for the development of the proposed methodology is also given. We then devote Section 3.3 to some estimation issues concerning the optimisation of the penalized log-likelihood and automatic smoothness selection. The methods are finally illustrated by fitting a bivariate system of equations to the levels of injury sustained by various parties involved in vehicle crashes (Section 3.4). Using data from the French ONISR, we compare several alternative scenarios and show: (i) how risk factors can have a peculiar influence on the same types of responses if different collision settings are considered; (ii) the various degrees of non-linearities characterising the effects of the continuous regressors; and (iii) the differences in the effects that risk factors have on the probability to sustain a certain injury severity level under several model specifications. Conclusions are drawn in Section 3.5.

3.2 Methods

We consider the pair of random variables $\boldsymbol{Y} := (Y_1, Y_2)^\top$ defined on the finite lattice $\mathcal{K}$ generated by the Cartesian product $\mathcal{K}_1 \times \mathcal{K}_2$, where $(\mathcal{K}_j, \preceq)$ is a totally ordered set for every $j \in \{1, 2\}$, and $\mathcal{K}_j := \{1, \ldots, K_j\}$ represents the levels of the categorical variable Y_j. The totality assumption implies that, under the binary relation $\preceq$, every element $k_j \in \mathcal{K}_j$ is comparable amongst all the others in the set. In the real data situations considered in this work, this excludes that a certain injury level cannot be appropriately recorded by the police officers. For instance, this may happen whenever a driver fails to be assigned to a pre-specified severity category after a vehicle accident had occurred.

In road safety studies, interest often lies in finding the risk factors associated to the specific injury severities incurred. Mathematically, this is achieved by investigating the effects that a given set of independent variables, as encoded in the array $\boldsymbol{x} := \text{vec}(\boldsymbol{x}_1, \boldsymbol{x}_2)$, have on some meaningful functions of the conditional joint mass of the random vector $\boldsymbol{Y}$. We consider therefore

$$r(\pi_k) := \mathbb{P}[Y_1 \preceq k_1, Y_2 \preceq k_2 | \boldsymbol{X} = \boldsymbol{x}] = \sum_{\tilde{k}_1 \preceq k_1} \sum_{\tilde{k}_2 \preceq k_2} \pi_{\tilde{k}_1, \tilde{k}_2}(\boldsymbol{x}), \tag{3.1}$$

where $k := (k_1, k_2)$. The vector $\boldsymbol{x}_j$ is assumed to collect the M_j explanatory variables of Y_j. Upon extending the approach introduced by Peyhardi et al. (2014) to multivariate categorical

responses, we define a copula regression for bivariate ordinal polychotomous outcomes as

$$\boldsymbol{r}(\bar{\boldsymbol{\pi}}) = \boldsymbol{g}^{-1}(\boldsymbol{\eta}) := (\boldsymbol{\mathcal{C}} \circ \boldsymbol{\mathcal{F}})(\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_{K-1}) \in [0,1]^{\#(\mathcal{K})-1}, \tag{3.2}$$

where

$$(\mathcal{C} \circ \mathcal{F})(\boldsymbol{\eta}_k) := \mathcal{C}_\gamma\big(F_{1,1}(\eta_{1,k_1}), F_{1,2}(\eta_{2,k_2}); \gamma\big) \in [0,1]$$

is the vector of the evaluations of link functions at the linear predictors $\boldsymbol{\eta}_k \in \mathbb{R}^2$ for every pair $k := (k_1, k_2)^\top \in \mathcal{K} \setminus \{K\}$, and $K := (K_1, K_2)^\top$. The linear predictors will be defined precisely in Section 3.2.2. We abide the convention of denoting the dimensionality of the distribution function with the first subscript of F.

The map $\mathcal{C}_\gamma : [0,1]^2 \longrightarrow [0,1]$ denotes a two-dimensional copula function, in which the dependence between the marginals $\{F_{1,j}(\eta_{j,k_j})\}_{j=1,2}$ is measured by an association parameter, γ, that binds them together. In general terms, a bivariate copula is a joint distribution function defined on $[0,1]^2$ with uniformly distributed marginals. An analytical definition of bivariate copulae is given precisely in Appendix B.2.1. In particular, if the $\{F_{1,j}\}$'s are univariate distribution functions, the Slkar's theorem ensures that the composition map $(\mathcal{C} \circ \mathcal{F})$ is indeed a 2-dimensional cdf with margins $F_{1,1}$ and $F_{1,2}$ (Sklar, 1959). We stress that the above representation is the bivariate equivalent of a Cumulative Link Model, whose various specifications are incorporated in (3.2) by appropriately characterising the distributions $F_{1,j}$'s. For instance, by fixing $F_{1,j} \equiv \Phi$ for each j, the Standard Normal cdf, and $\mathcal{C}_\rho$ the bivariate Gaussian copula with correlation ρ, the k-th element of (3.2) defines a bivariate ordered probit regression as

$$r(\pi_k) = \Phi_2(\Phi^{-1}(\Phi(\eta_{1,k_1})), \Phi^{-1}(\Phi(\eta_{2,k_2})); \rho) = \Phi_2(\eta_{1,k_1}, \eta_{2,k_2}; \rho).$$

As for any model involving ordinal responses, two restrictions complement the above representation. Let $\bar{k}$ and k be any two elements of $\mathcal{K}$ such that $\bar{k} \preceq k$ under a lexicographic order, then we require: (i) $r(\pi_{\bar{k}}) \leq r(\pi_k)$, and (ii) $\boldsymbol{\eta}_{\bar{k}} \leq \boldsymbol{\eta}_k$. The discussion of any formal argument concerning these constraints falls beyond the scope of this work, hence we refer the reader to Chapter 4 and references therein for a more thoughtful illustration of the issue. At this stage, it is just worthwhile to remark that the definition of some further parameters, termed cut points and denoted by c_{j,k_j}'s, allows us to account for (ii) simply by requiring $\{c_{j,k_j}\}_{k_j}$ to be an increasing sequence for every j. We also set $c_{j,K_j} = +\infty$ and $c_{j,1-1} =: c_{j,0} = -\infty$

in order to unbound the support of each linear predictor, hence becoming the extended real line $\mathbb{R} \cup \{-\infty, +\infty\}$.

We are now in the position to articulate (3.2) for any given level k of $\boldsymbol{Y}$ and available data $\mathbf{X} := \text{diag}(\mathbf{x}_1^\top, \mathbf{x}_2^\top)$ as

$$r(\pi_k) = (\mathcal{C}_\gamma \circ \mathcal{F})(\boldsymbol{c}_k - \mathbf{X}\boldsymbol{\beta}) = (\mathcal{C}_\gamma \circ \mathcal{F})(\mathbf{Z}\boldsymbol{\beta}_k), \tag{3.3}$$

where $\mathbf{Z}$ is the model design matrix with $\mathbf{Z} := \text{diag}(\mathbf{z}_1^\top, \mathbf{z}_2^\top)$ and $\mathbf{z}_j := (1, -\text{x}_{j,1}, \ldots, -\text{x}_{j,M_j})^\top$, and $\boldsymbol{\beta}_{j,k_j} := (c_{j,k_j}, \beta_{j,1}, \ldots, \beta_{j,M_j})^\top \in \mathbb{R}^{M_j+1}$ collecting the regression coefficients. Hence we have $\boldsymbol{\eta}_k = \mathbf{Z}\boldsymbol{\beta}_k \in \mathbb{R}^2$ for any pair k of $\mathcal{K}$.

3.2.1 The Class of Archimedean Copulae

Model (3.3) has been left intentionally generic, so that any copula function is in principle allowed to bind together the marginal distributions included in the model representation. However, the proposed implementation is practically restricted to the case of Standard Normal marginals and to the class of Archimedean copulae. This set up has some appealing features which makes its implementation particularly attractive. Specifically, for any given marginal distribution, Archimedean copulae are defined as the associative class of functions with generator $\psi : [0, 1] \longrightarrow [0, +\infty)$ which is assumed to be continuous, convex, decreasing, d-monotone, differentiable and such that $\psi(1) = 0$ (McNeil and Nešlehová, 2009); namely

$$\psi(\mathcal{C}_\gamma(F_{1,1}, F_{1,2})) = \psi(F_{1,1}) + \psi(F_{1,2}) = \psi(\Phi(\eta_{1,k_1})) + \psi(\Phi(\eta_{2,k_2})).$$

At this stage of the research, the marginals have been set to the Standard Normal cdf to ease the comparison with the relevant applied literature in which this distribution is widely used (see also Chapter 2). Nonetheless, representation (3.3) would in principle allow us to employ different marginals since the theoretical (and computational) framework remain essentially unchanged. Table 3.1 lists the various copula functions implemented in this work, whereas a graphical representation of their contours is provided in Appendix B.2. For completeness, we also consider rotated versions of the Clayton, Gumbel and Joe, which allow us to model negative dependences (for the 90 and 270 degrees), otherwise not implied by the respective canonical definitions. Analytically, rotations are computed using the definitions of Brechmann and Schepsmeier (2013) as reported in Appendix B.2; an example from the Joe copula is plotted in Figure 3.1. Two possible shortcomings may emerge from the above

Name	$\mathcal{C}_\gamma(u,v)$	Support of γ	γ^*
Gaussian	$\Phi_2(\Phi^{-1}(u),\Phi^{-1}(v))$	$[-1,1]$	$\tanh^{-1}(\gamma)$
Clayton	$(u^{-\gamma}+v^{-\gamma}-1)^{-1/\gamma}$	$(0,\infty)$	$\log(\gamma-\varepsilon)$
Frank	$-\gamma^{-1}\log[1+(e^{-\gamma u}-1)(e^{-\gamma v}-1)/(e^{-\gamma}-1)]$	$\mathbb{R}\setminus\{0\}$	$\gamma-\varepsilon$
Gumbel	$\exp\left\{-[(-\log u)+(-\log v)]^{1/\gamma}\right\}$	$[1,\infty)$	$\log(\gamma-1)$
Joe	$1-[(1-u)^\gamma+(1-v)^\gamma-(1-u)^\gamma(1-v)^\gamma]^{1/\gamma}$	$(1,\infty)$	$\log(\gamma-1-\varepsilon)$

Table 3.1: Families of some bivariate copula functions with association parameter γ. For optimisation purposes, an appropriate transformation γ^*, given in the last column of the table, is used in the estimation algorithm. The quantity ε denotes the machine smallest floating point multiplied by 10^6, and is introduced to force the transformed association parameters to lie in their respective supports throughout estimation. Finally, we have defined u and v to denote the marginals $\Phi(\eta_{j,k_j})$ for $j=1,2$.

specification of bivariate Archimedean copulae. A first one relates to their characterisation through a unique association parameter γ. Although restrictive, a possible alleviation can be achieved by defining a suitable transformation of γ (namely γ^* as in Table 3.1) in terms of a linear predictor too. This is effectively equivalent to assume a different association parameter for every individual in the sample. A further limitation is that the copulae employed in this chapter are exchangeable (Durante, 2009; Frees and Valdez, 1998). In the context of two-car accidents, for example, exchangeability implies that the probability of the two drivers to sustain a certain level of injury severities is invariant to weather $(Y_1=k_1,Y_2=k_2)$ or $(Y_2=k_2,Y_1=k_1)$ because they both give rise to the same bivariate distribution. For a complete account of copulae and their theoretical properties we refer the reader to the monograph of Nelsen (2006), whereas an excellent practical guide to copula modelling is offered by Trivedi and Zimmer (2005).

From an interpretational standpoint, equation (3.3) explicates that the cumulative mass function of the random vector $\boldsymbol{Y}$ is linked to the cdf of a latent (unobserved) random vector $\boldsymbol{Y}^*:=(Y_1^*,Y_2^*)^\top$, with support the extended real plane, through the equivalence

$$\{\boldsymbol{Y}\preceq k\}\Longleftrightarrow\{\boldsymbol{\varepsilon}\leq\boldsymbol{\eta}_k\},$$

where $\boldsymbol{\varepsilon}:=(\varepsilon_1,\varepsilon_2)^\top$ denotes the stochastic component in the regression of $\boldsymbol{Y}^*$ onto the columns of $\mathbf{X}$. In our context, we assume that each crash configuration carries a certain propensity to injury severities as described by the variables in $\mathbf{X}$ and the random components $\boldsymbol{\varepsilon}$. This propensity is then monotonically translated into an observed injury severity *level* sustained by the individuals involved in the accident. The construct above is standard in ordinal response modelling and traces back at least to the work of McKelvey and Zavoina (1975), with an explicit reference also in McCullagh (1980). The probability of the event

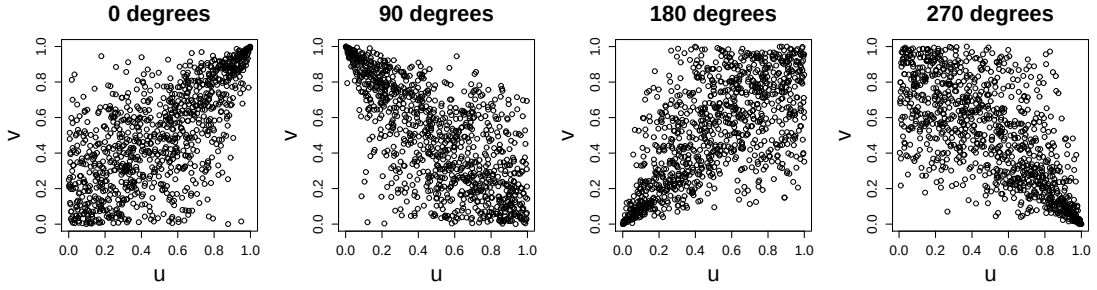

Figure 3.1: Random samples of 1,000 observations obtained from the two-place Joe copula with both Standard Normal marginals and different degrees of rotation. The association parameter has been fixed such that it corresponds to a Kendall's τ of 0.5 (-0.5 in case of 90 and 270 degrees). Explicit correspondences between γ and Kendall's τ in the context of Archimedean copulae are standard, and can be found in Brechmann and Schepsmeier (2013), for example.

$\{\boldsymbol{Y} = k\}$, namely π_k, is finally recovered by inverting the right-hand side of (3.3) which results in $\pi_k = r^{-1}(r(\pi_k)) = \sum_{l,m\in\{0,1\}}(-1)^{l+m} r(\pi_{k_1-l,k_2-m})$.

3.2.2 Penalized Regression Splines Approximation

Different covariate types are included in the proposed model. In particular, for any continuous regressor $\mathrm{v}_{j,l_j} \in \mathbb{R}$, like drivers' age or time of the accident (as expressed in hours and minutes), we advocate a non-parametric approach to curve estimation using penalized regression splines. That is, let their functional forms be smooth curves, $s_{j,l_j} : \mathbb{R} \longrightarrow \mathbb{R}$, then we seek to represent such covariate effects without the imposition of any given parametric structure. Let us assume first that we observe a sample of size n, $\{\mathbf{x}_i, \mathbf{y}_i\}_i$. Then, by appropriately choosing $H_j + 1$, $H_j := H(j, l_j) < n$, knot points in the interior of $[\mathrm{v}_{j,l_j,(1)}, \mathrm{v}_{j,l_j,(n)}]$, with $\mathrm{v}_{j,l_j,(i)} \leq \mathrm{v}_{j,l_j,(i+1)}$ for any i, it is possible to approximate the generic s_{j,l_j}-th curve as a linear combination of known basis spline functions, $\boldsymbol{b}_{j,l_j}$, and corresponding unknown coefficients, $\boldsymbol{\delta}_{j,l_j}$, to be estimated alongside the other model components. In other words we set

$$s_{j,l_j}(\mathrm{v}_{j,l_j,i}) \approx \boldsymbol{\delta}_{j,l_j}^{\top} \boldsymbol{b}_{j,l_j}(\mathrm{v}_{j,l_j,i}),$$

where the above vectors are H_j-dimensional. Moreover, since the curve estimates are only identified up to an intercept term, a centering constraint of the form $\mathbf{1}_n^{\top} \boldsymbol{s}_{j,l_j} = 0$, with $\boldsymbol{s}_{j,l_j}$ being the vector whose i-th element is $s_{j,l_j}(\mathrm{v}_{j,l_j,i})$, has to be imposed. This is achieved by employing the parsimonious method proposed by Wood (2006).

Basis functions are usually chosen based on their mathematical tractability and numerical stability. Among the most widely used in applications, we mention the cubic, pe-

nalized B-splines (Eilers and Marx, 1996) and thin plate regression splines (Wood, 2003), which are all supported by the computational routine attached to this chapter. Notice that the bases can be included in the design matrix by specifying the arrays $\mathbf{X}_{[j,l_j]} := (\boldsymbol{b}_{j,l_j}(\mathrm{v}_{j,l_j,1})|\cdots|\boldsymbol{b}_{j,l_j}(\mathrm{v}_{j,l_j,n}))^\top \in \mathbb{R}^{n\times H_j}$ and, accordingly, $\boldsymbol{\beta}_{[j,l_j]} := \boldsymbol{\delta}_{j,l_j} \in \mathbb{R}^{H_j}$. Thus it holds that

$$\boldsymbol{\eta}_j = \boldsymbol{c}_j - \mathbf{X}_{j,1}\boldsymbol{\beta}_{j,1} - \cdots - \mathbf{X}_{j,M_j}\boldsymbol{\beta}_{j,M_j} = \mathbf{Z}_j\boldsymbol{\beta}_j \in \mathbb{R}^n$$

is the linear predictor corresponding to the j-th equation, with $\mathbf{Z}_j := (\mathbb{I}_j, -\mathbf{X}_{j,1}, \ldots, -\mathbf{X}_{j,M_j})$ and $\boldsymbol{\beta}_j := \mathrm{vec}(\boldsymbol{c}_{j,k_j}, \boldsymbol{\beta}_{j,1}, \ldots, \boldsymbol{\beta}_{j,M_j})$, where $\mathbb{I}_j := \mathrm{diag}(\mathbb{1}_{\mathrm{y}_{j,i}=k_j})_{i,k_j} \in \{0,1\}^{n\times K_j-1}$ and $\boldsymbol{c}_{j,k_j} := (c_{j,k_j})_{k_j} \in \mathbb{R}^{K_j-1}$. The above representation is pivotal in applied modelling as it includes both non- and purely parametric covariate effects. In the statistical literature this form is commonly termed semi-parametric (e.g. Ruppert et al., 2003, Wood, 2006) and, once qualified with equation (3.2), the additive extension of a CLM emerges. We consequently label it a Cumulative Link Additive Model, which reads as

$$r(\boldsymbol{\pi}) = \mathcal{C}_2(\Phi(\mathbf{Z}_1\boldsymbol{\beta}_1), \Phi(\mathbf{Z}_2\boldsymbol{\beta}_2); \gamma) \in [0,1]^n,$$

where $\boldsymbol{\pi} := (\pi_1, \ldots, \pi_n)^\top [0,1]^n$ and $\pi_i := \mathbb{P}[\mathrm{y}_{1,i} = k_1, \mathrm{y}_{2,i} = k_2]$ for $i = 1, \ldots, n$.

3.2.3 Motivating the Proposed Bivariate Model

As pointed out by Abay et al. (2013), ideally all the people involved in the same car accident should be modelled simultaneously because all are affected by identical (or specular) crash conditions and occurrences. Ignoring this issue, for instance by pooling individuals together across all crashes and estimating individual-level injury severities, may lead to inefficient model parameter estimates as well as biased ones whenever unobserved heterogeneity is also present. We refer to Mannering and Bhat (2014) for an authoritative review of this and other related issues.

The structure of the model employed in this article is therefore similar to that of a bivariate system of SUR equations. The proposed representation accounts for the inter-relations between the injury severity levels sustained by any two people involved in the same vehicle collision (either one- or two-car crashes). This is achieved by estimating the impacts that some factors specific to the individuals (e.g., age, sex, seat position in the vehicle) have on the responses, as well as fitting a corresponding copula association parameter. The role of the latter is primarily to account for the influence that crash-specific unobservables may

have on the co-determination of the injury severities as reported by the police personnel. With respect to the study of multiple vehicle occupants, for example, such underlying factors may include variables like vehicle maintenance records, condition of safety equipments and presence or absence of some safety features. All of these are not always included in the official records.

3.3 Parameter Estimation

Under the usual *i.i.d.* conditions of the data generating process, the log-likelihood function corresponding to any bivariate model for ordinal responses representable in form (3.3) is given by

$$\ell(\boldsymbol{\vartheta}|\mathbf{y}_1, \mathbf{y}_2, \mathbf{X}_1, \mathbf{X}_2) = \sum_{i=1}^{n} \sum_{k \in \mathcal{K}} \mathbb{1}_{\mathrm{y}_{1,i}=k_1} \mathbb{1}_{\mathrm{y}_{2,i}=k_2} \log \pi_k(\mathbf{x}_{1,i}, \mathbf{x}_{2,i}),$$

where $\boldsymbol{\vartheta} := \mathrm{vec}(\boldsymbol{c}_1, \boldsymbol{c}_2, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \gamma)$ is the p-dimensional parameter vector, $\mathbf{x}_{j,i}^{\top}$ denotes the i-th row of $\mathbf{X}_j$, and the expression of the joint probability mass has been given in Section 3.2.1. The notation $\pi_k(\mathbf{x}_{1,i}, \mathbf{x}_{2,i})$ stresses the fact that, for each i, we observe a specific level $k \in \mathcal{K}$ of the vector of responses. Notice that the dependence of π_k on the available data is only through the systematic model components, $\eta_{j,i}$'s, which incorporate potentially any parametric and non-parametric effects of continuous covariates.

In order to make the model coherent in terms of its definition and parametrisation (cf. Section 3.2), we next transform some of the parameters and perform estimation over a modified vector $\widetilde{\boldsymbol{\vartheta}} \in \mathbb{R}^p$. However, to avoid clutter in the notation, from now on we will indifferently use $\boldsymbol{\vartheta}$ to denote both the transformed and the original parameter vector, as the distinction will be clearly inferred from the context. In particular, the setting of γ^* (as given in the last column of Table 3.1) is copula-specific and ensures that an unconstrained optimisation algorithm can be employed in the derivation of the Maximum Likelihood Estimator (MLE). The cut points are normalised through a squared polynomial transform, namely $\tilde{c}_{j,1} = c_{j,1}$ and $\tilde{c}_{j,k_j} := \sqrt{c_{j,k_j} - c_{j,k_j-1}}$ for any $k_j \in \mathcal{K}_j \setminus \{1\}$ and all j, so that $c_{j,k_j} = c_{j,k_j-1} + \tilde{c}_{j,k_j}^2 \geq c_{j,k_j-1}$. Alternative parametrisations are nonetheless available and have been proposed in the literature; for instance, in a similar context Kawakatsu and Largey (2009) used $c_{j,k_j}(\exp\{\tilde{c}_{j,k_j}\})$.

3.3.1 Penalized GLM Representation

Following the terminology of Peyhardi et al. (2014), model (3.3) gives the $(r, F_2, \mathbf{Z})$ representation of a GLM for categorical responses, with the additional feature that the foreseen bivariate distribution has to be now replaced by the copula function, namely $F_2 \equiv \mathcal{C}_\gamma(F_{1,1}, F_{1,2})$. Differently from a pure GLM, however, the non-parametric specification of the functional form of covariate effects and spatial variation within the data is likely to result in overfitting unless a suitable regularisation is introduced. To this end, a ridge-type penalisation acting on the elements of $\boldsymbol{\vartheta}$ is specified. That is, we associate the quadratic form $\mathcal{P}_{j,l_j} := \lambda_{j,l_j} \boldsymbol{\beta}_{j,l_j}^\top \overline{\mathbf{S}}_{j,l_j} \boldsymbol{\beta}_{j,l_j}$ to the (j, l_j)-th covariate, where $\overline{\mathbf{S}}_{j,l_j} = \mathbf{0}$ is assumed for parametric model components (i.e. no penalisation is attached to them). Furthermore, let $\lambda_{j,l_j} \in [0, +\infty)$ be a tuning parameter that controls for the trade-off between smoothness and fit: as $\lambda_{j,l_j} \to 0$ the estimated effects become wiggler and the fit perfect, and vice versa whenever $\lambda_{j,l_j} \to \infty$. The selection of the "right" amount of smoothness is therefore important in regression splines modelling and an appropriate method to deal with it is discussed in Section 3.3.3.

Upon setting $\boldsymbol{\lambda} := (\lambda_{j,l_j})_{l_j,j}$ and $\overline{\mathbf{S}}_{\boldsymbol{\lambda}} := \text{diag}(\lambda_{j,l_j} \overline{\mathbf{S}}_{j,l_j})_{l_j,j}$, we define the overall penalty $\mathbf{S}_{\boldsymbol{\lambda}}$ as $\overline{\mathbf{S}}_{\boldsymbol{\lambda}}$ padded with zeros such that $\mathcal{P}_{\boldsymbol{\lambda}} := \boldsymbol{\vartheta}^\top \mathbf{S}_{\boldsymbol{\lambda}} \boldsymbol{\vartheta} = \boldsymbol{\beta}^\top \overline{\mathbf{S}}_{\boldsymbol{\lambda}} \boldsymbol{\beta}$. In the following paragraphs we illustrate the two types of penalty matrices used to adapt our generic representation to the specific model components relevant to the case study considered in this work.

Continuous Covariates Regression splines account for non-linear smooth effects with varying degrees of complexity. The corresponding elements in $\boldsymbol{\lambda}$ are then associated to a conventional measure of curvature typically defined through an integrated square second derivative spline penalty. Namely, we set $\overline{\mathbf{S}}_{j,l_j} := \int \boldsymbol{b}''_{j,l_j} (\boldsymbol{b}''_{j,l_j})^\top \text{dv}_{j,l_j}$ with the integration conducted over the whole range of v_{j,l_j} (e.g. Green and Silverman, 1994).

Spatial Effects A location variable can be included in the model to control for the influence that some geographical-specific factors may have on the phenomenon under investigation. At the same time, it may also be the case that the resulting effects vary smoothly between nearby spatial regions. Consider a researcher interested in studying the consequences of fastening the seat belts (or wearing a helmet) on the severity of injuries in vehicle crashes. These predictors would arguably depend on the legal system of a country, but are also likely to be related to cultural attitudes and social sensibility of individuals. These may in turn be affected by the presence of a local community effect of some sort. Therefore, it

is reasonable to assume that this influence is spatially-dependent, instead of exhausting its effects just outside the specific geographical unit or community it refers to.

A random Markov field (RMF) smoother is implemented to exploit the spatial information in the data, and is suitable whenever a given area is made up of discrete contiguous units, as the 96 Departments of continental France. A RMF is commonly regarded as a generalisation of a univariate first order random walk to two dimensions. To illustrate this idea, let us assume that we have R regions indexed by r, so that the spatial covariate effect of the i-th regressor is given by $\mathbf{x}_{r_i}^\top \boldsymbol{\beta}_r$, with $\boldsymbol{\beta}_r := (\beta_{r,1}, \ldots, \beta_{r,R})^\top$. The corresponding design matrix is constructed such that its (i,r)-th element equals 1 if observation i belongs to r, and 0 otherwise. The underlying assumption of the construction is that neighbourhood sites are more alike than two arbitrary ones. In particular, any two sites, r and s, are said to be neighbour if they share at least a common boundary; we denote by δ_r the set of regions adjacent to r, and $N_r := \#(\delta_r)$ its cardinality. Following Kneib (2005), we next assign a prior probability to the evaluation of the spatial smoother corresponding to each region of the form

$$\beta_{r,r'} | \beta_{r,s}, s \neq r', \sigma_\beta^2 \sim \mathcal{N}\left(\frac{1}{N_{r'}} \sum_{s \in \delta_{r'}} \beta_{r,s}, \frac{\sigma_\beta^2}{N_s} \right) \quad r', s = 1, \ldots, R.$$

Thus the conditional mean of $\beta_{r,r'}$ is described as the unweighted average of the function evaluations of all neighbour sites. Finally, for any two regions r and s, the penalty matrix associated to the spatial covariate can be shown to be given by the adjacency matrix (Rue and Held, 2005)

$$\overline{\mathbf{S}}_{[r,s]} := \begin{cases} -1 & r \neq s \ \wedge \ s \in \delta_r \\ 0 & r \neq s \ \wedge \ s \notin \delta_r \\ N_r & r = s \end{cases} .$$

3.3.2 Estimating $\boldsymbol{\vartheta}$ Given the Smoothing Parameters

Generalized Linear Models in the $(r, F_2, \mathbf{Z})$ form augmented by a regularisation term $\mathcal{P}_{\boldsymbol{\lambda}}$ can be estimated within a penalized likelihood (PL) framework. The corresponding MPLE is then defined as the solution of the problem

$$\widehat{\boldsymbol{\vartheta}} := \arg\max_{\boldsymbol{\vartheta}} \left\{ \ell(\boldsymbol{\vartheta}|\cdot) - \frac{1}{2} \boldsymbol{\vartheta}^\top \mathbf{S}_{\boldsymbol{\lambda}} \boldsymbol{\vartheta} \right\}, \tag{3.4}$$

which is optimised for any given value of the smoothing parameter vector $\boldsymbol{\lambda}$. Notice that the simultaneous optimisation of the above criterion with respect to $(\boldsymbol{\vartheta}, \boldsymbol{\lambda})$ would clearly

result in over-fitted estimates. In this case, in fact, any solution would satisfy the points $\widehat{\lambda}_{j,l_j} = 0$ for all l_j and any j. Therefore, by letting $\ell_{\mathrm{p}}(\boldsymbol{\vartheta}, \boldsymbol{\lambda}|\cdot)$ be the penalized log-likelihood corresponding to the argument of (3.4), we seek to estimate model parameters by means of the outer iteration algorithm introduced by O'Sullivan et al. (1986). Specifically, the approach consists of iteratively maximising $\ell_{\mathrm{p}}(\boldsymbol{\vartheta}^{[\alpha]}|\boldsymbol{\lambda}^{[\alpha]}, \cdot)$ with respect to $\boldsymbol{\vartheta}$ first, and then employing the estimates so produced to obtain updated smoothing parameters. The two steps are then repeated until convergence.

In practice, (3.4) is maximised using a trust-region algorithm which is generally more stable and faster than its line-search counterparts, especially for functions that exhibit non-linearities or regions that are close to flat (Nocedal and Wright, 2006, Ch. 4). Other optimisation algorithms are of course allowed and implementable. For the $[\alpha]$-th iteration, the routine solves the sub-problem

$$\begin{aligned} \min_{\boldsymbol{p}} \ \widetilde{\ell}_{\mathrm{p}} &= \min_{\boldsymbol{p}} \left\{ - \left[\ell_{\mathrm{p}}(\boldsymbol{\vartheta}^{[\alpha]}) + \boldsymbol{p}^{\top} \nabla_{\boldsymbol{\vartheta}^{[\alpha]}} \ell_{\mathrm{p}}(\boldsymbol{\vartheta}^{[\alpha]}) + \tfrac{1}{2} \boldsymbol{p}^{\top} \nabla_{\boldsymbol{\vartheta}^{[\alpha]} \boldsymbol{\vartheta}^{[\alpha]\top}} \ell_{\mathrm{p}}(\boldsymbol{\vartheta}^{[\alpha]}) \boldsymbol{p} \right] \right\} \Big|_{\boldsymbol{\lambda} = \boldsymbol{\lambda}^{[\alpha]}} \\ &\quad \text{subject to } \|\boldsymbol{p}\| \leq \Delta^{[\alpha]} \\ \\ \boldsymbol{\vartheta}^{[\alpha+1]} &= \boldsymbol{\vartheta}^{[\alpha]} + \boldsymbol{p}^{[\alpha+1]}, \end{aligned} \tag{3.5}$$

where $\boldsymbol{p}^{[\alpha+1]} := \arg\min_{\boldsymbol{p}} \widetilde{\ell}_{\mathrm{p}}(\boldsymbol{\vartheta}^{[\alpha]}|\boldsymbol{\lambda}^{[\alpha]}, \cdot)$. The above defines a quadratic approximation of the negative log-likelihood about $\boldsymbol{\vartheta}^{[\alpha]}$. This problem is then used to choose the best step $\boldsymbol{p}^{[\alpha+1]}$ within the trust-region, namely the ball centered in $\boldsymbol{\vartheta}^{[\alpha]}$ of radius $\Delta^{[\alpha]}$. The main advantage of this algorithm compared to its line-search counterparts is its superiority in terms of computational time and stability. Whenever the objective function is undefined or indeterminate at the solution of sub-problem (3.5), that proposal is rejected, the trust-region shrunken, and the optimisation re-stated accordingly. In other words, every step $(\boldsymbol{\vartheta}^{[\alpha+1]}|\boldsymbol{\lambda}^{[\alpha]})$ gives always a solution for the trust-region problem. Details on the numerical routine as employed in our estimation scheme are given in Geyer (2013), which also discusses stability issues and termination criteria.

Analytic derivations of the score and Hessian employed in the computation of (3.5) are obtained by exploiting the multivariate GLM structure of model (3.3) and follow the construction given in Section 2.3. In particular, by letting $\ell_{\mathrm{p},i}(\boldsymbol{\vartheta}|\cdot)$ be the penalized log-likelihood contribution of the i-th observation, we obtain

$$\nabla_{\boldsymbol{\vartheta}} \ell_{\mathrm{p},i}(\boldsymbol{\vartheta}) = \frac{\partial \boldsymbol{\eta}_k}{\partial \boldsymbol{\vartheta}} \left(\frac{1}{\pi_k} \frac{\partial \boldsymbol{F}_k}{\partial \boldsymbol{\eta}_k} \frac{\partial \mathcal{C}_\gamma}{\partial \boldsymbol{F}_k} \frac{\partial \pi_k}{\partial \boldsymbol{r}_k} \right) - \mathbf{S}_{\boldsymbol{\lambda},i} \boldsymbol{\vartheta} = \mathbf{D}_i^{\top} \mathbf{u}_i - \mathbf{S}_{\boldsymbol{\lambda},i} \boldsymbol{\vartheta} =: \boldsymbol{g}_{\mathrm{p},i},$$

with $\mathbf{D}_i^\top := (\partial \boldsymbol{\eta}_k / \partial \boldsymbol{\vartheta})$, and

$$\nabla_{\boldsymbol{\vartheta}\boldsymbol{\vartheta}^\top} \ell_{\mathrm{p},i}(\boldsymbol{\vartheta}) = \mathbf{D}_i^\top \left[\frac{1}{\pi_k} \left\{ \frac{\partial \boldsymbol{F}_k}{\partial \boldsymbol{\eta}_k} \frac{\partial^2 \mathcal{C}_\gamma}{\partial \boldsymbol{F}_k \partial \boldsymbol{F}_k^\top} \left(\frac{\partial \boldsymbol{F}_k}{\partial \boldsymbol{\eta}_k} \right)^\top + \frac{\partial^2 \boldsymbol{F}_k}{\partial \boldsymbol{\eta}_k \partial \boldsymbol{\eta}_k^\top} \frac{\partial \mathcal{C}_\gamma}{\partial \boldsymbol{F}_k} \right\} \frac{\partial \pi_k}{\partial \boldsymbol{r}_k} - \mathbf{u}_i \mathbf{u}_i^\top \right] \mathbf{D}_i - \mathbf{S}_{\boldsymbol{\lambda},i},$$

where the term in the square brackets $[\cdot]$ of the equation above is commonly labelled $\mathbf{W}_i \in \mathbb{R}^{5\times 5}$, and is the multivariate analogous of the weight matrix in classical iterative GLM estimation. Notice that, upon defining

$$\boldsymbol{r}_k := (r(\pi_{k_1-1,k_2-1}), r(\pi_{k_1-1,k_2}), r(\pi_{k_1,k_2-1}), r(\pi_{k_1,k_2}))^\top \in [0,1]^4,$$

it holds that $\partial \pi_k / \partial \boldsymbol{r}_k = (1,-1,-1,1)^\top$ because of the way the mass function is recovered from (3.1), whilst $\mathbf{D}_i$ is of dimension $(5 \times p)$ and includes the derivatives of the cut points and of the covariate vector. The analytical score and Hessian are implemented in `CopulaCLM` and they have been verified using numerical differentiation. Finally we set the quantities $\mathbf{D} := (\mathbf{D}_1^\top | \cdots | \mathbf{D}_n^\top)^\top$, $\mathbf{u} := \mathrm{vec}(\mathbf{u}_1, \ldots, \mathbf{u}_n)$ and $\mathbf{W} := -\mathrm{diag}(\mathbf{W}_1, \ldots, \mathbf{W}_n)$ to obtain the global expressions needed for the algorithm's development.

3.3.3 Estimating $\boldsymbol{\lambda}$ given $\boldsymbol{\vartheta}$

Given the multidimensional nature of our framework, computations may become burdensome if a direct grid search optimisation of, for instance, a prediction error criterion is used to smoothness selection. It is therefore essential to be able to estimate $\boldsymbol{\lambda}$ in an automatic way. To this end, we adopt a modified version of the Un-biased Risk Estimator recently applied by Marra et al. (2015) in the context of bivariate dichotomous response models. The key idea is to base the derivation of a penalized iterative re-weighted least squares estimator on the Hessian (or Fisher Information) matrix and score vector considered globally rather than the single components that make them up (e.g. Yee and Wild, 1996). Traditional methods involve the computation of the square root and inversion of $\mathbf{W}$, which are typically more likely to fail to be positive definite for a subset of observations.

To overcome this inconvenience, let us define $\overline{\mathbf{W}} := \mathbb{E}[\mathbf{W}]$, and $\mathcal{I} := \mathbf{D}^\top \overline{\mathbf{W}} \mathbf{D}$ be the Fisher information of the un-penalized log-likelihood; it holds that $\mathbf{W} = \overline{\mathbf{W}} + o_p(1)$. Then, by computing a first-order Taylor expansion of $\boldsymbol{g}_\mathrm{p}$ about the MPLE, and appropriately re-arranging its terms, we get an iterative algorithm of the form

$$\boldsymbol{\vartheta}^{[\alpha+1]} = (\mathcal{I}^{[\alpha]} + \mathbf{S}_{\boldsymbol{\lambda}}|_{\boldsymbol{\lambda}=\boldsymbol{\lambda}^{[\alpha]}})^{-1} \sqrt{\mathcal{I}^{[\alpha]}} \mathbf{z}^{[\alpha]},$$

where $\mathbb{R}^p \ni \mathbf{z}^{[\alpha]} := \sqrt{\mathcal{I}^{[\alpha]}}\boldsymbol{\vartheta}^{[\alpha]} + \sqrt{\mathcal{I}^{[\alpha]}}^{-1}\boldsymbol{g}^{[\alpha]}$ is the pseudo-data vector associated to the penalized GLM model. Notice that, from likelihood theory, asymptotically it holds that $\mathbf{z} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{I}_p)$, where $\boldsymbol{\mu} := \sqrt{\mathcal{I}}\boldsymbol{\vartheta}_0$ is evaluated at the true parameter vector. Define now $\widehat{\boldsymbol{\mu}}$ the plug-in estimator obtained from the Generalized Least Squares estimate $\widehat{\boldsymbol{\vartheta}}_{\text{GLS}}$ of the regression of $\mathbf{z}$ onto the columns of $\mathcal{I}$ and ridge penalty $\mathcal{P}_{\boldsymbol{\lambda}}$. Namely, $\widehat{\boldsymbol{\mu}} := \sqrt{\mathcal{I}}\widehat{\boldsymbol{\vartheta}}_{\text{GLS}} = \mathbf{P}_{\boldsymbol{\lambda}}\mathbf{z}$, with $\mathbf{P}_{\boldsymbol{\lambda}} := \sqrt{\mathcal{I}}(\mathcal{I} + \mathbf{S}_{\boldsymbol{\lambda}})^{-1}\sqrt{\mathcal{I}}$ being the corresponding model influence matrix. Then we seek to estimate $\boldsymbol{\lambda}$ in such a way that the resulting non-parametric covariate effects are as close as possible to the real ones, that is by suppressing any complex structure which is not supported by the available data. To this end, we compute

$$\mathbb{E}\|\boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}\|^2 = \mathbb{E}\|\mathbf{z} - \mathbf{P}_{\boldsymbol{\lambda}}\mathbf{z}\|^2 - \widetilde{n} + 2\text{tr}(\mathbf{P}_{\boldsymbol{\lambda}}), \tag{3.6}$$

where $\widetilde{n} := 5n$ and $\text{tr}(\mathbf{P}_{\boldsymbol{\lambda}})$ defines the effective degrees of freedom (edf) of the penalized model. Hence an estimator for the smoothing parameters is defined iteratively as

$$\begin{aligned}\boldsymbol{\lambda}^{[\alpha+1]}|\boldsymbol{\vartheta}^{[\alpha+1]} &:= \arg\min_{\boldsymbol{\lambda}} \mathcal{V}(\boldsymbol{\lambda}) \\ &:= \arg\min_{\boldsymbol{\lambda}} \left\{ \|\mathbf{z}^{[\alpha+1]} - \mathbf{P}_{\boldsymbol{\lambda}}|_{\boldsymbol{\lambda}=\boldsymbol{\lambda}^{[\alpha]}}\mathbf{z}^{[\alpha+1]}\|^2 - \widetilde{n} + 2\text{tr}(\mathbf{P}_{\boldsymbol{\lambda}})|_{\boldsymbol{\lambda}=\boldsymbol{\lambda}^{[\alpha]}} \right\},\end{aligned}$$

which is implemented using the stable and efficient routine of Wood (2004). A detailed derivation of equivalence (3.6) can be found in Chapter 2 in an analogous setting and with similar notation. The two steps detailed in this and previous sections are iterated until the stopping criterion is satisfied: $\max|\boldsymbol{\vartheta}^{[\alpha+1]} - \boldsymbol{\vartheta}^{[\alpha]}| < 10^{-6}$. They can be summarised as follows:

1. for a given value of the parameter vector $\boldsymbol{\vartheta}^{[\alpha]}$, and holding $\boldsymbol{\lambda}^{[\alpha]}$ fixed, estimate:

$$\boldsymbol{\vartheta}^{[\alpha+1]} = \arg\min_{\boldsymbol{p}} \widetilde{\ell}_{\text{p}}(\boldsymbol{\vartheta}^{[\alpha]}|\boldsymbol{\lambda}^{[\alpha]}, \cdot) + \boldsymbol{\vartheta}^{[\alpha]};$$

2. construct the pseudo-data vector $\mathbf{z}^{[\alpha+1]}$ and $\mathbf{P}_{\boldsymbol{\lambda}}$ using $\boldsymbol{\vartheta}^{[\alpha+1]}$ and find an estimate of $\boldsymbol{\lambda}$:

$$\boldsymbol{\lambda}^{[\alpha+1]} = \arg\min_{\boldsymbol{\lambda}} \mathcal{V}(\boldsymbol{\lambda}).$$

It is worthwhile to remark that the above scheme is implemented only from the knowledge of score and Hessian. In principle, therefore, any likelihood model regularised by a ridge-type penalisation term can be naturally extended to the various types of covariate effects

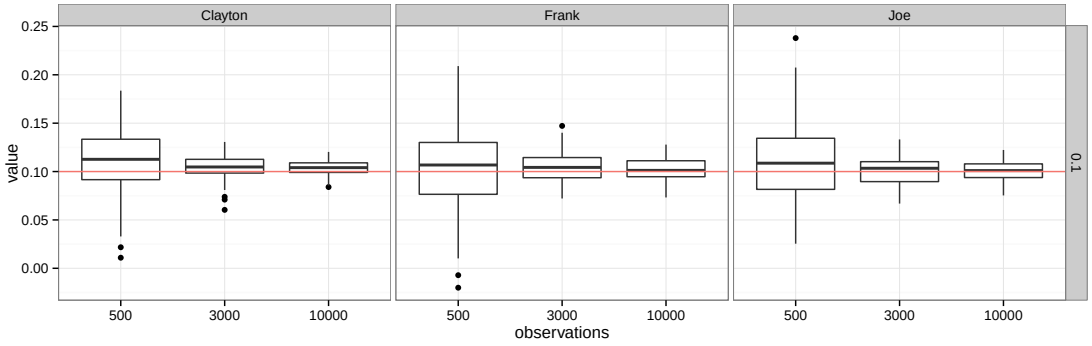

Figure 3.2: Box plots corresponding to the estimates of the copula association parameter γ for different sample sizes and copulae employed. The coefficient γ has been reported under its corresponding Kendall's τ correlation, whose true simulated value ($\tau = 0.1$) is depicted as the red line in each panel. Results are obtained from 100 replications of the DGP detailed in Appendix B.2.

described in this article in a relatively straightforward way. Potentially, also beyond the pure GLM family.

3.3.4 Some Simulation Evidence

A small set of Monte Carlo experiments has been conducted to investigate the finite-sample properties of the MPLE for the proposed copula model. For the sake of space, the exact definitions of the data generating processes (DGPs) employed are given in Appendix B.2, so we just comment here on the results obtained. Figure 3.2 shows the behaviour of the estimates of the copula association parameter under different settings and sample sizes. Specifically, the association parameter is reported using the equivalent Kendall's τ metric to facilitate the comparison between the various copula scenarios. Its magnitude has been set to match that found in the real-data application. In line with expectations, we report that, as the sample size increases, the proposed MPLE approaches the true value with a lower standard deviation, and the parameter of interest is recovered reasonably well also at "modest" n of around $3,000$. This is a remarkable finding considering that a low magnitude of the association between the two equations locates γ close to the lower bound of its support (or to zero in the case of the Frank), which may result in numerical instabilities as well as make it hard to detect a dependence (Trivedi and Zimmer, 2005).

Some evidence of the ability of our model to recover the non-parametric covariate effects is provided in Figure 3.3, in which three smooth functions were included in the DGP: two referring to the equation for Y_1^*, and one to that for Y_2^*. The curves recovered at each replication illustrate graphically the degree of uncertainty attached to smooth function esti-

mation, a concept formalised in Marra and Wood (2012) for the construction of point-wise Bayesian credible intervals in GAMs. Of course, less precise results are expected when fewer observations are used.

3.4 Data Analysis

Our empirical study uses data from the "Bulletins d'Analyse des Accidents Corporels" (BAAC) 2014. This dataset collects information about all vehicle accidents occurred in France that required the intervention of the police personnel. Agents were responsible for recording crash details, which were then centrally administrated by the ONIRS and subsequently published in the BAAC. At present, this comprises 4 headings referring to different accident features, and labelled "caractéristique", "lieux", "véhicules" and "usagers". Every accident is identified by a unique progressive serial number, identical for each vehicle and individual involved in the same crash.

Since the original dataset contains details on every kind of accident with at least one vehicle affected, we consider only those instances conforming with the features of primary interest for the present study. Accordingly, we select one-car crashes of four-wheels motor vehicles with two occupants (Scenario I), and two-vehicle collisions in which the injury severities of the two drivers are jointly modelled (Scenario II). The resulting datasets finally include $1,232$ and $20,079$ observations, respectively, and are available to download from the authors' website. Although some insights into the factors that influence injury outcomes can be drawn from univariate models for crash, vehicle or roadway types, the results obtained in this way may not be directly applicable to all traffic crash scenarios (Russo et al., 2014). Therefore, it may be the case that even identical risk factors can affect the same responses peculiarly, once different crash dynamics are considered. Hence it is important to discern these issues in various settings.

The aim of this analysis is to quantify the influence that some measurable risk factors of interest have on the probability that vehicles' occupants sustain a certain level of injury severity, while accounting for the possible presence of unobserved variables affecting their inter-relationships. A bivariate copula Cumulative Link Additive Model is estimated for this purpose. Specifically, injury severities are recorded by data collectors into four ordered categories: "no injury" (level 1), "non hospitalised" and "hospitalised" injuries (levels 2 and 3), and "fatal" (level 4) ones: a summary is reported in Table 3.2.

Scenario I					
Injury Severity Driver	Injury Severity Passenger				Marginals
	no injury	non hospitalised	hospitalised	fatal	
no injury	0 (0.00%)	188 (15.26%)	166 (13.47%)	17 (1.38%)	371 (30.11%)
non hospitalised	59 (4.79%)	251 (20.37%)	75 (6.09%)	10 (0.81%)	395 (32.07%)
hospitalised	69 (5.60%)	68 (5.52%)	213 (17.29%)	42 (3.41%)	392 (31.82%)
fatal	12 (0.97%)	11 (0.89%)	36 (2.92%)	15 (1.22%)	74 (6.00%)
Marginals	140 (11.36%)	518 (42.05%)	490 (39.77%)	84 (6.82%)	1,232 (100.00%)
Scenario II					
Injury Severity Driver A	Injury Severity Driver B				Marginals
	no injury	non hospitalised	hospitalised	fatal	
no injury	1,389 (6.92%)	6,190 (30.83%)	2,809 (13.99%)	138 (0.69%)	10,526 (52.43%)
non hospitalised	3,825 (19.05%)	1,348 (6.71%)	473 (2.36%)	49 (0.24%)	5,695 (28.36%)
hospitalised	2,206 (10.99%)	582 (2.90%)	613 (3.05%)	68 (0.34%)	3,469 (17.28%)
fatal	165 (0.82%)	90 (0.45%)	107 (0.53%)	27 (0.13%)	389 (1.93%)
Marginals	7,585 (37.78%)	8,210 (40.89%)	4,002 (19.93%)	282 (1.40%)	20,079 (100.00%)

Table 3.2: Distributions of injury severities sustained by driver and passenger (Scenario I) and by the two drivers (Scenario II) in vehicle-related accidents obtained using BAAC 2014 data. The categorisation follows the information recorded by the police personnels on the place of crash.

3.4.1 Models and Results

The proposed model specification follows closely previous works, with the aim of facilitating the comparison of results (e.g. Yamamoto and Shankar, 2004, Eluru et al., 2010 for Scenario I; or Russo et al., 2014 for Scenario II). Specifically, in line with the BAAC organisation, we group the explanatory variables into four macro-areas referring to occupant (driver or passenger), vehicle, motorway and accident characteristics: they are all listed in the first column of Tables B.5 and 3.4.

The continuous covariates included in the analysis, namely age of the occupants and time of collision, are estimated non-parametrically and approximated using penalized thin plate regression splines. Moreover, since both of these variables are usually categorised to achieve a different fit for each corresponding ordinal level (e.g. in Eluru et al., 2010 and Chiou et al., 2013), our methodology contributes to the road safety literature by providing the researcher with an automatic data-driven way to modelling flexibly covariate effects. A better model fit

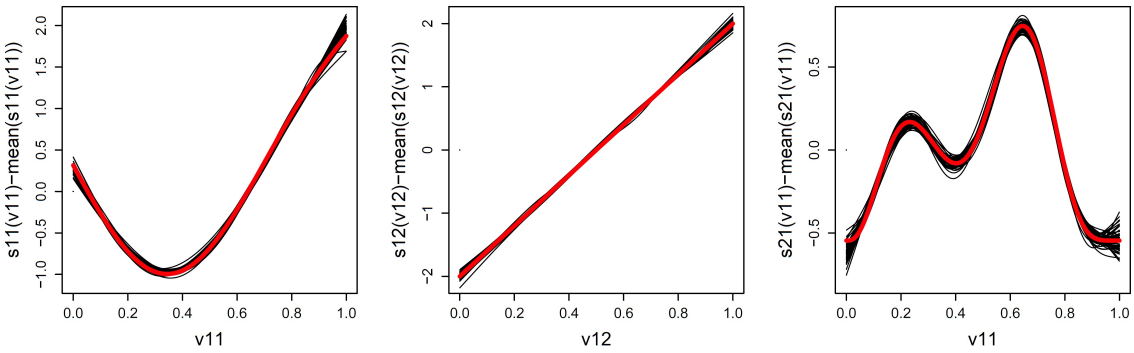

Figure 3.3: Estimated smooth curves obtained from 50 replicates of a Monte Carlo experiment comprising 10,000 simulated observations of a Joe copula model (true curves reported in red). The smooth components were represented using penalized thin plate regression splines with basis dimensions equal to 10 and penalties based on second-order derivatives. Results are plotted on the scale of the linear predictors.

is therefore expected in this case. A regional variable identifying the location of each crash has also been included in the model. This aims at estimating Department-specific factors at a lower geographical level than France as a whole (see for instance Figures 3.5 and 3.6).

Estimation has been performed using the authors' `R` function `CopulaCLM` for all the copula specifications reported in Table 3.1, alongside with their corresponding degrees of rotation. Thus, a total of 15 different models were fitted, including the one referring to the Independent copula obtained by pooling together all the observations and by estimating a univariate ordered probit regression. Based on the Bayesian Information Criterion (BIC), as reported in Table 3.3, the preferred models are those based on the Joe_0 in both Scenario I and II. Interestingly, the supported copulae are all those defined for a positive, albeit low, association parameter γ. Moreover, we notice that the preferred models exhibit an upper tail dependence, indicating that extreme values of the propensity to severity injuries for one individual tend to be associated with extremes recorded for the other individual in the study. For illustration purposes, all the results that follow are computed by employing the Joe_0 copula distribution.

As previously discussed in Section 3.2.3, in our context the copula association parameter measures the influence that common unobserved factors may have on the sustainment of certain injury levels of road users. On this point, previous research has generally reported that a positive association exists between the two equations in the system. This means that, on average, unobservables tend to drive in the same direction the injury severities sustained by the occupants in the same motor vehicle in one-car accidents, and by both drivers in two-car crashes. Overall, we stress that every model not supported by an information criterion

Copula	Scenario I			Scenario II		
	$\widehat{\gamma}$	(se)	BIC	$\widehat{\gamma}$	(se)	BIC
Independent	–	–	6,253.07	–	–	–
Gaussian	0.0118	(0.0354)	6,262.80	0.0580	(0.0119)	73,697.18
Clayton_0	0.0000	(–)	6,258.98	0.0000	(–)	74,403.08
Clayton_{90}	0.0110	(–)	6,258.29	0.0000	(∞)	73,729.99
Clayton_{180}	0.2281	(0.0489)	**6,224.60**	0.1450	(0.0099)	**73,415.78**
Clayton_{270}	0.0000	(–)	6,258.99	0.0000	(∞)	73,730.03
Frank	0.1945	(0.2115)	6,261.00	0.0330	(0.0816)	73,726.24
Gumbel_0	1.0780	(–)	6,238.46	1.0662	(0.0055)	73,419.85
Gumbel_{90}	1.0000	(–)	6,258.91	1.0000	(–)	73,727.57
Gumbel_{180}	1.0000	(–)	6,258.99	1.0000	(–)	73,725.38
Gumbel_{270}	1.0000	(–)	6,259.01	1.0050	(0.0972)	73,758.50
Joe_0	1.2130	(0.0463)	**6,219.17**	1.1056	(0.0078)	**73,327.50**
Joe_{90}	1.0000	(–)	6,253.06	1.0000	(–)	73,663.97
Joe_{180}	1.0000	(–)	6,259.04	1.0000	(–)	73,723.00
Joe_{270}	1.0000	(–)	6,259.06	1.0108	(0.0168)	73,734.78

Table 3.3: Estimated association parameters for the different copula models considered in the chapter, with corresponding standard errors reported in brackets. The last column shows the associated BIC, with the selected models highlighted in bold. Since the penalty matrix in the estimation algorithm can suppress some dimensions of the parameter space, we have: $\text{BIC} = -2\ell_{\text{p}}(\widehat{\boldsymbol{\vartheta}}|\cdot) + edf \log n$, where edf are the estimated degrees of freedom as defined in Section 3.3.3. Notice that, wherever the algorithm did not converge, the standard errors were not reported. The BIC for the independent case in Scenario I is computed on one parameter less than the others, while the one of Scenario II is not given because based on double the number of observations, and so misleading. Similar results were obtained when the Akaike Information Criterion was used. Standard errors are obtained by simulation from the posterior distribution of the MPLE; details on the scheme are in Section 2.3.4.

either failed to converge, or indicated the absence of any association among the two equations. This behaviour is not uncommon in copula modelling (Trivedi and Zimmer, 2005), and it has been reported also for bivariate systems of equations with dichotomous responses by Radice et al. (2015) and Marra et al. (2015), for instance. We are unaware, however, whether such a pattern has been recognised in the road safety literature too, since we find this information neither disclosed nor discussed. It would be nonetheless interesting to further investigate this issue in empirical studies, especially in those settings experiencing various strengths of the copula association parameter. This would give thoughtful insights on a source of model mis-specification stemming from the incompatibility between the observed data structure and the restrictions imposed on the support of γ.

Estimated Effects In Table 3.4 the estimates corresponding to the coefficients of the purely parametric covariate effects are reported for Scenario II, alongside with those arising from the independent model. Results from Scenario I are instead detailed in Appendix B.2 (Table B.5). Notice that, by pooling all the individuals together, in Scenario II the number

of observations doubled compared to the employment of a bivariate system of equations. For Scenario I, instead, we preferred to employ a bivariate Gaussian model with γ structurally constrained to zero as independence benchmark. We believe that drivers and passengers are often subject to different effects for the same risk factor, and this would in turn distort any analysis conducted on a single pooled model. Consider, for example, different types of collisions. Qualitatively, the estimates obtained show that a sideswipe collision to the left would result in a higher severity injury propensity for the driver; however, whenever it occurs to the right, the passenger is likely to be the most affected party. This distinction is nonetheless smoothed and not captured in a single-equation model (see Table B.6 in Appendix B.2 for the whole set of estimates referring to the pooled independent model for Scenario I).

Overall, the sign of many of the estimated coefficients confirms the results previously reported in the accident literature, and are consistent with expectations. In particular, for Scenario II, we find that females show a higher propensity to sustain severe injuries when compared to males, regardless of the vehicle they are actually seated in. This gender difference might be related to weight, body mass and other factors, and is in line with other authors' findings (see, for instance, Ulfarsson and Mannering, 2013). Also, travelling in four-wheels motor vehicles is generally associated with lower injuries than is travelling on motorcycles, with larger ones ($> 125\text{cm}^3$) being unsafer as compared to smaller motorised two-wheelers ($< 125\text{cm}^3$). These results are intuitive since a better protection from severe injuries can be expected in cars, whereas small motorcycles may be constrained to reduced speed limits and restricted circulation on faster roadways. Among environmental factors, lighting in force at night lowers the propensity of severe injury; however an opposite sign is found for the driver and passenger (refer to Scenario I for this effect, Table B.5). Similar contrasting effects emerge also when we compare adverse weather conditions against normal ones in both scenarios. These results are quite surprising: if on the one hand the presence of water or ice on street pavement may reduce vehicles' friction, hence fostering the likelihood of an accident, on the other hand cars may proceed at a reduced speed and drivers be more cautious (Eluru et al., 2010). Analogous arguments can be made in those instances of reduced visibility, like foggy weather conditions. In any case, systematic differences between driver and passenger, as well as among Scenarios I and II, strengthen the assertion that risk factors act differently on sustained injury severities in various crash settings, and thus any generalisation of the results has to be carefully assessed. Some implications for roadway

Scenario II: Estimates						
Variables	Driver A		Driver B		Independent model	
	estimates	(se)	estimates	(se)	estimates	(se)
Occupant Characteristics						
Gender (male)						
female	0.2624	(0.0197)	0.3393	(0.0187)	0.3030	(0.0136)
Vehicle Characteristics						
Type (motorcycle $< 125\text{cm}^3$)						
Motorcycle $> 125\text{cm}^3$	0.1145	(0.0303)	0.1197	(0.0257)	0.1235	(0.0198)
Vehicle M1	−1.6171	(0.0221)	−1.4841	(0.0211)	−1.5058	(0.0151)
Vehicle N1	−1.8863	(0.0445)	−1.8396	(0.0435)	−1.7876	(0.0307)
Motorway Characteristics						
Intersection (off intersection)						
X	−0.0207	(0.0230)	−0.0392	(0.0218)	−0.0418	(0.0158)
T	−0.2531	(0.0276)	−0.1420	(0.0256)	−0.2053	(0.0188)
Y	−0.2296	(0.0659)	−0.2046	(0.0622)	−0.2274	(0.0455)
> 4 branches	−0.2339	(0.0804)	−0.3008	(0.0745)	−0.2753	(0.0550)
roundabout	−0.4326	(0.0555)	−0.3119	(0.0505)	−0.3970	(0.0374)
circus/square	−0.3288	(0.0911)	−0.1795	(0.0762)	−0.2299	(0.0579)
level crossing	0.4659	(0.4589)	0.3154	(0.4564)	0.3719	(0.3259)
other	0.0323	(0.0782)	0.0647	(0.0743)	0.0537	(0.0539)
Type (motorway)						
Route Nationale	0.1127	(0.0456)	0.2653	(0.0448)	0.2096	(0.0321)
Route Départementale	0.1137	(0.0308)	0.3302	(0.0302)	0.2385	(0.0216)
Voie Communale	−0.4268	(0.0306)	−0.1146	(0.0294)	−0.2626	(0.0213)
off public road network	−0.1714	(0.3129)	−0.9176	(0.3734)	−0.5129	(0.2399)
parking	−0.4512	(0.1652)	−0.1439	(0.1521)	−0.3182	(0.1128)
other	−0.1409	(0.0999)	−0.0637	(0.0974)	−0.0993	(0.0702)
Circulation regime (missing)						
one-way	−0.1292	(0.0407)	−0.2092	(0.0385)	−0.1695	(0.0280)
two-way	0.0441	(0.0347)	−0.0414	(0.0330)	0.0068	(0.0240)
presence of median	−0.1611	(0.0432)	−0.1620	(0.0413)	−0.1764	(0.0299)
other	−0.0914	(0.1140)	−0.1813	(0.1098)	−0.1293	(0.0796)
Horizontal alignment (straight)						
left curve	0.2276	(0.0332)	0.1626	(0.0324)	0.2037	(0.0232)
right curve	0.3766	(0.0329)	0.1360	(0.0327)	0.2734	(0.0232)
S	0.4602	(0.0799)	0.1906	(0.0804)	0.3394	(0.0568)
Accident Characteristics						
Lighting (daylight)						
sunrise/sunset	0.0382	(0.0374)	0.0662	(0.0359)	0.0281	(0.0268)
night without street lights	0.4835	(0.0366)	0.4634	(0.0363)	0.4198	(0.0272)
night, street lights in force	−0.0960	(0.0291)	−0.0151	(0.0272)	−0.1143	(0.0221)
Atmospheric conditions (normal)						
light rain	0.0804	(0.0270)	0.0367	(0.0258)	0.0665	(0.0187)
heavy rain	0.2539	(0.0543)	0.0664	(0.0535)	0.1804	(0.0382)
snow	0.5011	(0.2025)	0.7263	(0.1971)	0.6305	(0.1417)
fog	0.2952	(0.1120)	0.2818	(0.1116)	0.2836	(0.0791)
heavy wind/storm	0.2765	(0.2283)	−0.0297	(0.2282)	0.1117	(0.1613)
clear	0.0998	(0.0829)	0.2796	(0.0797)	0.1842	(0.0573)
clouds	0.2150	(0.0463)	0.1227	(0.0453)	0.1770	(0.0324)
Manner of collision (missing/other)						
head-on	0.0521	(0.0340)	0.0841	(0.032)	0.0861	(0.0236)
rear-end	−0.1068	(0.0470)	−0.0864	(0.0370)	−0.0363	(0.0287)
sideswipe, right	0.0589	(0.0480)	−0.0811	(0.0478)	−0.0076	(0.0341)
sideswipe, left	0.0868	(0.0487)	0.0141	(0.0432)	0.0618	(0.0326)
Passenger (no)						
yes	0.0744	(0.0244)	0.0051	(0.0210)	0.0702	(0.0159)
Security device (not put on)						
put on	−0.0499	(0.0233)	−0.0661	(0.0240)	−0.0294	(0.0168)
$c_{j,1}$	−1.2304	(0.0664)	−1.3065	(0.0640)	−1.1815	(0.0462)
$c_{j,2}$	−0.1244	(0.0057)	0.1221	(0.0051)	0.0746	(0.0038)
$c_{j,3}$	1.3644	(0.0097)	1.7942	(0.0100)	1.6557	(0.0071)
No. observations	20,079		20,079		40,158	

Table 3.4: Estimates and associated standard errors obtained for the parametric model components by applying `CopulaCLM` to the BAAC 2014 data in Scenario II when the Joe_0 copula is used.

design can be drawn too. Specifically, our analysis shows that the presence of roundabouts is likely to be associated with a reduction in injury propensity whenever two vehicles hit each other (Scenario II), whereas a corresponding increase emerges for both driver and passenger if only one car is involved in the crash (Scenario I). This is presumably a result of the implicit accident dynamics leading to the two scenarios, where the latter is likely to be mostly caused by a loss of control of the vehicle at entry, on the circulatory roadway or its exit. This suggests undertaking suitable actions in terms of design and safety countermeasures of roundabouts to make them an even more effective tool for prevention.

The estimated non-parametric model components for Scenario II are depicted in Figure 3.4, whereas those referring to Scenario I are reported in Appendix B.2 for the sake of space. Analogous conclusions may nonetheless be drawn, although the estimated effects are smoother, and both occupant's age and time of the accident appear in this case not to be important determinants of the driver's injury severity. This may be due to the relatively low sample size available for this scenario. The curves are estimated by low-rank penalized thin plate regression splines with basis dimension equal to 10 and penalties as described in Section 3.3.1. In line with the literature, our analysis highlights an almost steady effect of age up to around 40-45 years and it increases rapidly for people older than 60 years. This evidence deserves some attention: with an increasing number of elderly people in Europe, the implementation of ad-hoc actions and/or legislations seem to us of growing importance to foster road safety measures. The maps in Figure 3.4 depict the magnitude of the resulting estimates for each Department.

It is worthwhile stressing that parametric and non-parametric estimates have to be interpreted qualitatively, as they affect directly the propensity of injury, Y_J^*, rather than the responses as measured on their manifest ordinal scale. This practical limitation is accounted here by computing the model (pseudo-)elasticities, roughly interpretable as the percent change in the probability that the average individual sustains a certain injury level for 1% increase in a measured continuous covariate (e.g. Mannering, 2009). Specifically, for any categorical regressor and under the maintained assumptions, the pseudo-elasticity of the (j, m_j)-th covariate on the j-th response for individual i is defined as

$$\widehat{\mathcal{E}}_{\mathrm{x}_{j,m_j,i}}^{\mathbb{P}[\mathrm{y}_{j,i}=k_j]} := \sum_{l\in\{0,1\}} (-1)^l \left[\frac{\Phi(\eta_{j,k_j,i} - \beta_{j,m_j}\{\mathbb{1}_{l=0} - \mathrm{x}_{j,m_j,i}\}) - \Phi(\eta_{j,k_j-1,i} - \beta_{j,m_j}\{\mathbb{1}_{l=0} - \mathrm{x}_{j,m_j,i}\})}{\Phi(\eta_{j,k_j,i}) - \Phi(\eta_{j,k_j-1,i})} \right]_{\boldsymbol{\vartheta}=\widehat{\boldsymbol{\vartheta}}},$$

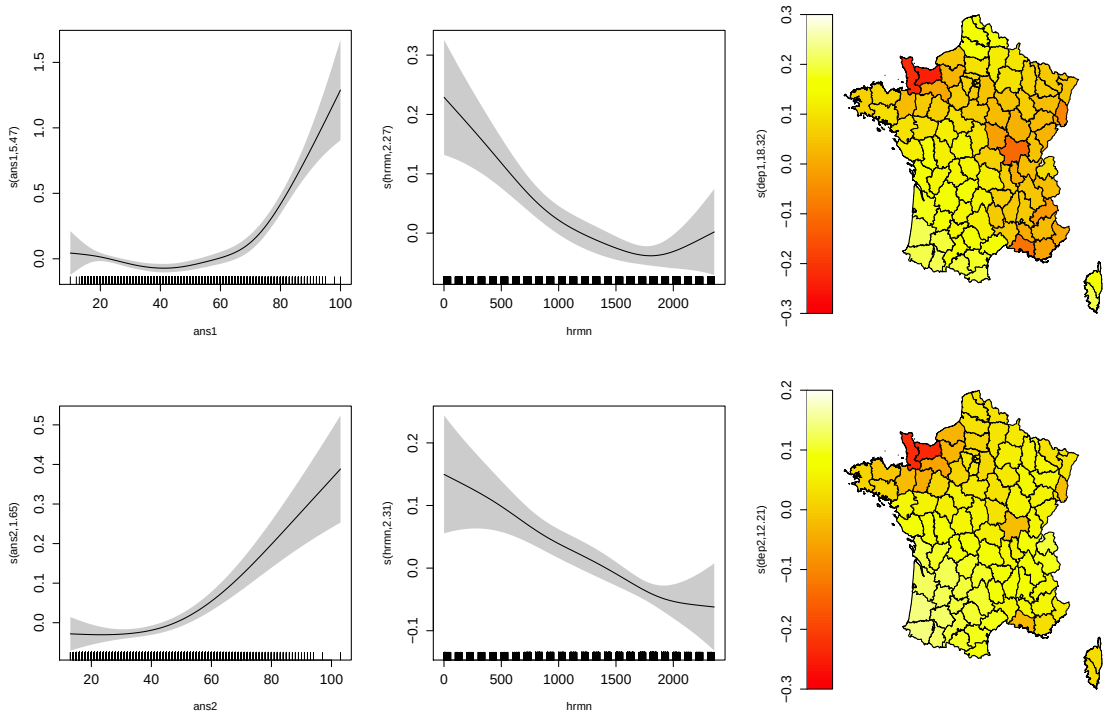

Figure 3.4: Smooth functions estimates and associated 95% point-wise confidence intervals corresponding to the two equations (first and second row) of the bivariate model applied to the BAAC 2014 data under Scenario II when using the Joe_0 error dependence. The curves relate to the effects of age and `time` (expressed in hours and minutes, `hrmm`) on the propensity of injury severities of drivers in 2-car collisions. Confidence intervals are based on the results of Marra and Wood (2012) for GAMs, which are accommodated into a bivariate penalized GLMs admitting a $(r, F_2, \mathbf{Z})$ representation as explained in Chapter 2. The effective degrees of freedom are reported into brackets in the y-axis caption, with a value of one corresponding to a straight line estimate. The covariate values are represented by a jittered rug plot at the bottom of each graph. The maps, instead, depict graphically the strength of the estimates obtained for the regional variable in each of the 96 Department of continental France.

which is averaged to obtain

$$\widehat{\mathcal{E}}^{\mathbb{P}[\mathrm{y}_j=k_j]}_{\mathrm{x}_{j,m_j}}(\widehat{\boldsymbol{\vartheta}}) = \left[\frac{1}{n}\sum_{i=1}^{n} \widehat{\mathcal{E}}^{\mathbb{P}[\mathrm{y}_{j,i}=k_j]}_{\mathrm{x}_{j,m_j,i}}(\boldsymbol{\vartheta})|_{\boldsymbol{\vartheta}=\widehat{\boldsymbol{\vartheta}}}\right] \cdot 100.$$

The statistic above estimates the percent change corresponding to the observation of a certain level of the (j, m_j)-th categorical covariate. For instance, it measures by how much (in percentage points) the probability of an individual being hospitalised changes (on average) when the crash occurs at roundabouts, with respect to all the other intersection types. Tables 3.5 and B.7 report all pseudo-elasticities for a number of competing models: they all confirm the previous considerations based on the models' estimates.

Figure 3.5 illustrates the roundabout effects computed at the geographical location of

Variables	SCENARIO II: PSEUDO-ELASTICITIES				
	Joe_0: Semi-parametric		Independent	Joe_0: Parametric	
	Driver A	Driver B	Drivers	Driver A	Driver B
	Occupant Characteristics				
Gender (male)					
female	51.2199	63.3263	57.2809	49.0147	64.3432
	Vehicle Characteristics				
Type (motorcycle $< 125cm^3$)					
Motorcycle $> 125cm^3$	21.5703	19.5992	22.4733	12.4107	17.7820
Vehicle M1	−95.1798	−93.4637	−93.9468	−95.6780	−94.1561
Vehicle N1	−97.3239	−97.0778	−96.7766	−97.7803	−97.4369
	Motorway Characteristics				
Intersection (off intersection)					
X	−3.3212	−5.8193	−6.3512	−1.4909	−3.9816
T	−31.7989	−19.7918	−28.0645	−32.4045	−18.1672
Y	−34.4756	−27.4212	−30.6469	−25.8402	−22.0858
> 4 branches	−32.2960	−37.9463	−35.9644	−19.6491	−26.0959
roundabout	−52.0617	−39.0754	−47.9295	−53.4887	−41.6091
circus/square	−42.5045	−24.4390	−30.9286	−31.5714	−7.8534
level crossing	104.3385	58.0468	73.3202	80.1874	42.7814
other	5.3802	10.2499	8.6798	−0.1093	5.8727
Type (motorway)					
Route Nationale	19.8356	47.4023	37.3954	12.6677	46.3187
Route Départementale	20.0250	61.3016	43.3151	12.3073	56.9322
Voie Communale	−51.5674	−16.2490	−34.5895	−52.6100	−15.1491
off public road network	−24.7185	−79.2843	−57.4700	−22.6495	−77.0557
parking	−53.6121	−20.0222	−40.4252	−58.2970	−25.1451
other	−20.7573	−9.3221	−14.5495	−29.4835	−13.7900
Circulation regime (missing)					
one-way	−19.1954	−27.9574	−23.7186	−7.2882	−19.7631
two-way	7.4037	−6.1352	1.0663	10.3137	−5.5498
presence of median	−23.4015	−22.2966	−24.5684	−13.2578	−16.0509
other	−13.9355	−24.6545	−18.5694	−12.4101	−24.4588
Horizontal alignment (straight)					
left curve	43.3808	27.3313	36.1979	36.2971	22.4362
right curve	79.4569	22.4980	50.7559	72.8857	17.1688
S	144.4867	36.7612	83.0723	125.7304	28.2268
	Accident Characteristics				
Lighting (daylight)					
sunrise/sunset	6.3774	10.4918	4.4718	9.3324	11.1976
night without street lights	109.5614	92.8252	85.1674	112.6790	88.6743
night, street lights in force	−13.9091	−2.2535	−15.6351	−3.8771	2.9762
Atmospheric conditions (normal)					
light rain	13.8325	5.7126	10.8377	14.0210	7.5702
heavy rain	49.2787	10.5419	31.5915	46.1942	7.8947
snow	114.8808	167.3511	144.4313	73.2345	130.8153
fog	58.9479	50.8455	52.9707	40.6242	39.3954
heavy wind/storm	54.5058	−4.4395	18.7279	41.6901	−11.5470
clear	17.3984	50.3945	32.3513	10.8402	41.9703
clouds	40.6199	20.1327	30.9299	39.4989	21.8518
Manner of collision (missing/other)					
heads-on	9.0503	3.4770	14.9423	7.3696	13.5057
rear-end	−15.2807	−12.4695	−5.4242	−16.4820	−9.6178
sideswipe, right	10.3250	−11.2449	−1.1826	6.9893	−10.1128
sideswipe, left	15.7586	2.1764	10.3965	13.2137	3.5122
Passenger (no)					
yes	12.7516	0.7857	11.4579	11.6577	−2.6750
Security device (not put on)					
put on	−7.6349	−9.3241	−4.4262	−10.0084	−9.5552
No. observations	20,079		40,158	20,079	

Table 3.5: Pseudo-elasticities of the parametric model components of Scenario II obtained by applying the preferred Joe_0 copula, independent and the purely parametric models. Quantities computed with respect to the hospitalised injuries.

each collision. We then compare them against the corresponding pseudo-elasticities obtained from the independent models. Two points are worth noticing. First of all, we report that not accounting for unobservables in the study may lead to overestimated elasticity effects, which may in turn distort the information transmitted to policymakers throughout their decision process. This is clearly seen in the bottom row for Scenario II, where a *univariate* model is compared against the preferred copula specification. The same figure referring to Scenario I shows instead that this difference is qualitatively almost indistinguishable. This may be explained by the use in the latter case of a *bivariate* independent model. In fact, since this assumes the product of two Normals as the reference distribution, as γ tends to the infimum of its support, the copula models converge indeed to the corresponding independence benchmark. On the other hand, this clarifies the pivotal role of multivariate modelling in multi-party vehicle collisions, even regardless of the actual strength of the common unobserved factors.

As a second point of interest, the different magnitude of these effects seems to be clustered between Departments, hence suggesting a possible location-dependent copula association parameter. Although this feature is not currently implemented in `CopulaCLM`, it can be achieved by specifying the parameter γ as a function of some relevant predictors, which may include a regional variable too. In road safety studies, Eluru et al. (2010) employed an analogous approach to accommodate the potential heterogeneity of people driving specific vehicle types. This generalisation may constitute a promising path for future extensions of the proposed model, also in light of different application fields.

The gains of applying a bivariate copula Cumulative Link Additive Model are perhaps better summarised in Figure 3.6, where we compare the pseudo-elasticities of S curves on hospitalised injuries obtained under alternative models. The Gaussian copula corresponds to a semi-parametric bivariate probit regression, analogous to that of Hillmann et al. (2014), in which model randomness is induced by the bivariate Standard Normal distribution. This assumption is precisely what we aimed at extending with the introduction of Archimedian copulae. Also, in the purely parametric model, drivers' age and time of collision are assumed to affect the responses linearly, as typically done in the applied road safety literature. In both cases, these traditional models tend to overestimate the effects computed by the copula model. Although these considerations may be less vivid for different covariates, we think that the proposed methodology can nonetheless constitute a valid way for researchers to test their estimates against different assumptions and scenarios.

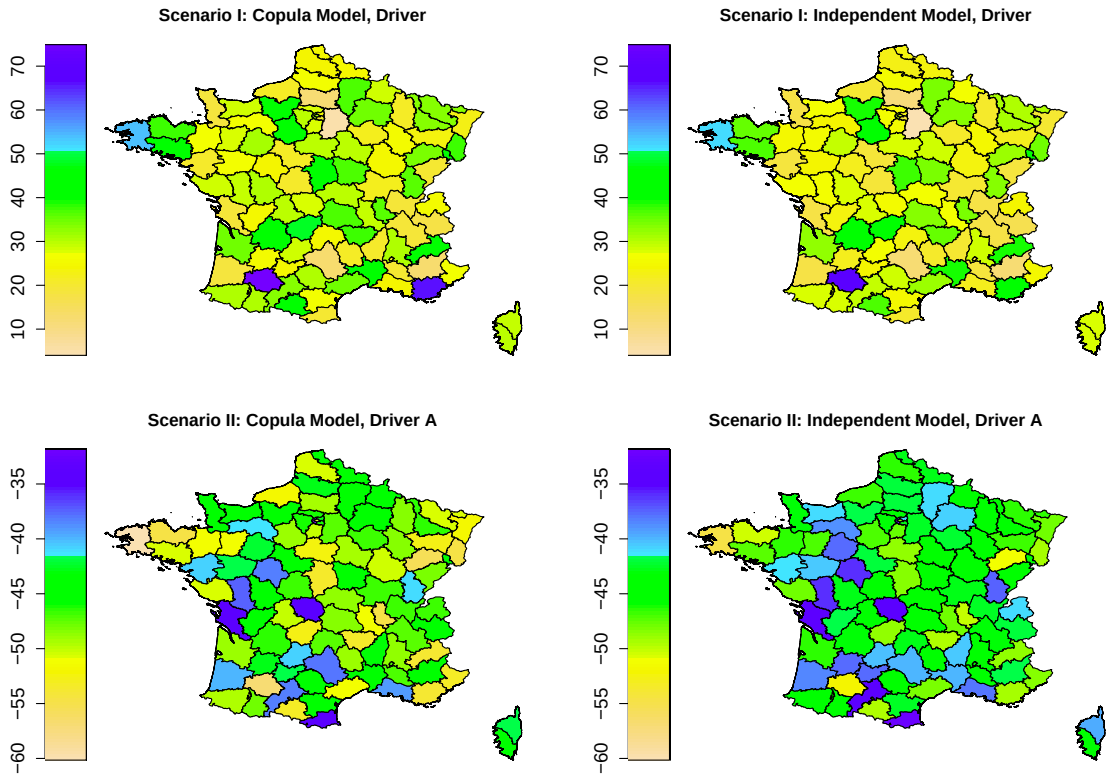

Figure 3.5: Pseudo-elasticities of the presence of roundabouts on the probability of the average occupant to sustain a hospitalised injury in the 96 French Departments. The comparison between copula and independent models is presented for Scenario I, top row, and Scenario II, bottom row. Notice: (i) the qualitative analysis of the coefficients' signs is enhanced by the formal computation of the (pseudo-)elasticities; and (ii) the difference in the estimates obtained when a pooled univariate model is employed rather than a bivariate one (Scenario II). In this case, only results for Driver A have been reported.

3.5 Discussion

This chapter introduced a flexible bivariate model for ordinal responses. The proposed framework can account for the presence of common unobserved variables influencing the inter-relationships between the outcomes, non-linear covariate effects and non-Gaussian dependences. The model has been described in terms of a copula-based additive extension of the classical Cumulative Link Model, whose representation has taken advantage of a penalized GLM form. The estimation algorithm has been discussed and all the necessary computational procedures incorporated in the freely-available routine `CopulaCLM` for the `R` environment.

The proposed methodology has been motivated by the study of two different scenarios within the road safety literature. Here, a bivariate specification is introduced to model jointly the injury severities sustained by individuals in vehicle accidents. Specifically, we

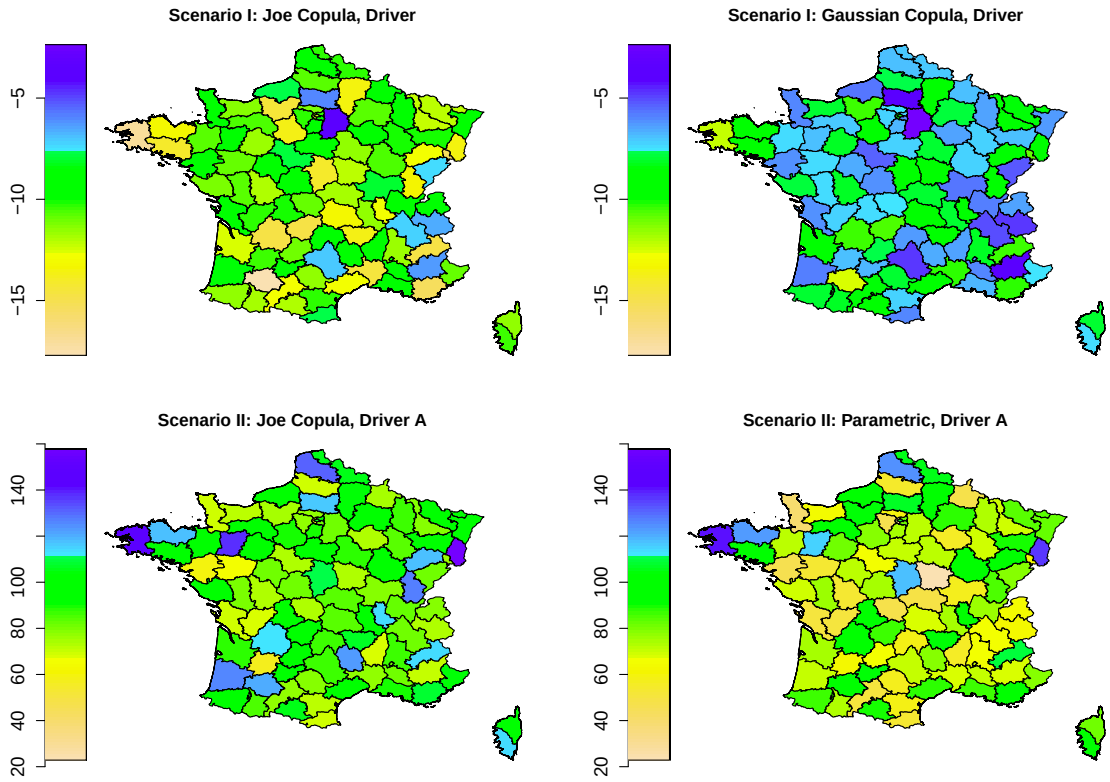

Figure 3.6: Some potential risks of model mis-specification: comparison between pseudo-elasticities of S curves on hospitalised injuries. The Gaussian copula (not preferred based on the BIC) corresponds to the use of a semi-parametric bivariate ordered probit regression. The parametric model, instead, neglects both non-linearities and smoothed variation in the regional variable documented in Figure 3.4. This highlights the need of using flexible models reducing the risk of mis-specification. Notice that less vivid results may be obtained when the effects of different covariates are computed: the whole results have been listed in Tables B.7 and 3.5.

dealt with one-car crashes of four-wheel vehicles with one passenger, and with collisions of two motorised vehicles. Using BIC, we chose the Joe_0 copula to represent the dependence between the two equations in the model, and a positive, though small, association has been documented. The effects that the induced dependence has on the model interpretation have been summarised in Figure 3.5 with respect to roundabouts. In the bottom row, in particular, we have reported the differences a researcher would have found if all the individuals involved in the same crash were not modelled separately, but analysed within the same univariate model.

The pseudo-elasticities estimated from the selected Joe_0 were plotted and compared against those from the Gaussian copula and a purely parametric model (Figure 3.6). The Gaussian model overestimates the effects at departmental level by an average of -3.26 (with minimal and maximal difference of -5.10 and -1.47, respectively). That is, under a

bivariate Gaussian assumption, the presence of an S curve reduces the probability of facing a hospitalised injury (resulting from a one-car collision in the average Department) of 3.26% more than what is estimated under a Joe_0. In the same vein, we found that a fully parametric model estimates that S curves increase the probability of hospitalisations by 14.21 percentage points less than a semi-parametric model does.

As previously remarked, an interesting avenue for future research would be to express the copula association parameter as a function of some relevant covariates to capture the effects of unobservables which may possibly be location- or vehicle type-dependent. The use of penalized regression splines in this setting can further enhance the flexibility of the model. Moreover, since dichotomous responses can be thought as a special case of discrete variables with only two levels, allowing one dependent variable to be binary will not change the essence of our theoretical and computational framework. This is a relevant point in the road safety literature since a possible source of endogeneity has been recognised when estimating the effects of fastening the seat belts on the sustained levels of injury in car accidents (e.g. de Lapparent, 2008). From a methodological point of view, this is an instance of unmeasured confounding that can be dealt with by setting up and estimating a bivariate recursive system of equations, for example in the spirit of Marra and Radice (2011).

Chapter 4

Discrete Responses in Generalized Additive Models

This chapter introduces a conceptual framework for the analysis of dichotomous and ordinal polychotomous responses within a penalized multivariate Generalized Linear Model. The proposed structure allows for a rather flexible predictor specification through the inclusion of non-parametric and spatial covariate effects, as well as the characterisation of the distribution of the stochastic model components with copulae of univariate marginals. The framework is subsequently illustrated through a non-random sample selection problem concerning the estimation of the HIV prevalence in Zambia using the 2007 DHS dataset.

Caution: The writing of this chapter required a non-negligible reduction of my sleeping time. The effects are clearly reflected on the writing style.

4.1 Introduction

Generalized Linear Models (GLMs, Nelder and Wedderburn, 1972) are a comprehensive class of models that allows researchers to conduct estimation and inference for a variety of response types within the same coherent unifying framework. However, despite their undoubted relevance in applied research, they rely on a purely parametric specification of the covariate effects on the response which effectively constraints the linear predictors to be of a determined fixed-order polynomial, for instance. This is a strong requirement, as one cannot typically expect to know in advance the actual form of covariate-response relationships: its incorrect specification would potentially generate a non-negligible source of bias.

An existing approach to overcome this limitation is to consider a more flexible class of models that permits the representation and estimation of the additive effects of some continuous covariates in a data-driven way. Methods of this kind are usually termed semi-parametric in the statistical literature because they conjugate both a parametric and a non-parametric characterisation of the functional forms of the regressors. Specifically, whenever the baseline structure is that of a GLM, the so-called Generalized Additive Models emerge.

They usually complement their parametric counterparts using a regression spline approach (Hastie and Tibshirani, 1986, 1990). Nonetheless, as any traditional regression analysis, GAMs are effectively models for the expected value of a random variable described by a certain conditional distribution function. To enhance flexibility, therefore, it is also desirable to extend the framework to qualify the dependence of any moment of order higher than one on some explanatory variables of interest. In this way, the risk of mis-specification is further alleviated. This approach usually comes under the name of distributional regression, whose ideas have been variously incorporated within a GAM setting. For example, Rigby and Stasinopoulos (2005) proposed a Generalized Additive Model for Location, Scale and Shape (GAMLSS), whose structure has been recently extended to a multivariate setting by Klein et al. (2015). A review of these and of some other existing methodologies is presented in Kneib (2013). This line of research seeks to achieve an even higher degree of flexibility by increasing the number of distributions allowed by the framework, and including in their respective specifications several kind of covariate effects.

The present work builds on these ideas in the context of discrete outcomes. Starting from the definition of a GAM for a J-variate vector of categorical responses, we discuss the conceptual representation of dichotomous and ordinal polychotomous dependent variables in terms of a triplet $(r, F_J, \mathbf{Z})$, and a penalty matrix $\mathbf{S}_{\boldsymbol{\lambda}}$ that allows for the use of linear predictors incorporating non-parametric, spatial and random covariate effects. A method for dealing with a mixture of those two types of responses is also outlined. We then show how a generic estimation algorithm can be derived, and inference subsequently conducted, within the resulting multivariate Generalized Additive Model. We finally argue that such algorithm can be, *mutatis mutandis*, applied to any model representable in the $(r, F_J, \mathbf{Z})$ form. Although the pace of the discussion is intentionally kept at a quite generic level, connections between the proposed framework and some existing models are made. These have the dual scope of motivating our representation with well-developed examples from the literature and, at the same time, offering a way to extend them to the more flexible predictor specifications that form the domain of our work. In particular, attention is given to nested models accounting for unmeasured residual confounding, an instance rather frequent in observational studies with detrimental consequences on the parameter estimates if not adequately controlled for (e.g. Becher, 1992). The proposed representation is then employed to define a non-random sample selection model for dichotomous responses to credibly assess the human immunodeficiency virus (HIV) prevalence in Zambia. With this empirical illus-

tration, we give evidence of the flexibility of our generic representation, which also allows for the inclusion of bivariate distributions defined through copulae of univariate marginals, and the dependence of the association parameter expressible as a functional of the available data.

4.2 A Penalized GLM Representation for Discrete Responses

Let $\boldsymbol{Y} = (Y_1, \ldots, Y_J)^\top$ be a random vector with support the discrete set $\mathcal{K} := \mathcal{K}_1 \times \cdots \times \mathcal{K}_J$, where $\mathcal{K}_j := \{1, \ldots, K_j\}$ and $\#(\mathcal{K}_j) = K_j < \infty$ for every $j \in \mathcal{J}$, $\mathcal{J} := \{1, \ldots, J\}$. We say that each variable Y_j has finite K_j levels. The set $\mathcal{K}_j$ is assumed to collect both qualitative and quantitative elements as well as variables measured on the nominal or ordinal scale (Stevens, 1946). Specifically, the former differentiates items based only on the categories they belong to, whereas the latter allows also for a rank order by which the realisations of Y_j can be sorted. However, the relative degree of difference between the levels lacks of any meaningful interpretation. For notational convenience, we represent each $k_j \in \mathcal{K}_j$ by a natural number and, wherever the support of Y_j is ordinal, we assume that the set $(\mathcal{K}_j, \preceq)$ is totally ordered under the binary relation $\preceq$.

In analogy with the approach outlined in Peyhardi et al. (2014) for the univariate case, we consider a regression of the probability $\pi_k = \mathbb{P}[\boldsymbol{Y} = k | \boldsymbol{X} = \boldsymbol{x}]$, with $k := (k_1, \ldots, k_J)^\top \in \mathcal{K}$, on some covariates $\boldsymbol{x} := \mathrm{vec}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_J)$ defined through a Generalized Linear Model form

$$\bar{\boldsymbol{\pi}} = \boldsymbol{g}^{-1}(\boldsymbol{\eta}) := (\boldsymbol{r}^{-1} \circ \boldsymbol{\mathcal{F}})(\boldsymbol{\eta}_1(\boldsymbol{x}_1), \ldots, \boldsymbol{\eta}_{K-1}(\boldsymbol{x}_{K-1})) \in \mathcal{M}, \tag{4.1}$$

where $\boldsymbol{r} : \mathcal{M} \longrightarrow \mathcal{P}$ is a diffeomorphism from $\mathcal{M} := \{(0,1)^{K-1} | \mathbf{1}^\top \bar{\boldsymbol{\pi}} < 1\}$ to an open subset $\mathcal{P}$ of $(0,1)^{K-1}$, with $K := \#(\mathcal{K})$. We also define the map $\boldsymbol{\mathcal{F}} : \mathcal{S} \longrightarrow \mathcal{M}$, with $\mathcal{S} \subset \mathbb{R}^{J \times (K-1)}$, and set accordingly

$$\begin{array}{rcll} \bar{\boldsymbol{\pi}} & := & (\pi_1, \ldots, \pi_{K-1})^\top & \in \ (0,1)^{K-1} \\ \boldsymbol{\eta} & := & (\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_{K-1}) & \in \ \mathbb{R}^{J \times (K-1)} \\ \boldsymbol{\mathcal{F}}(\boldsymbol{\eta}) & := & (F_J(\boldsymbol{\eta}_1), \ldots, F_J(\boldsymbol{\eta}_{K-1}))^\top & \in \ (0,1)^{K-1} \end{array}$$

where the last array collects fully-specified J-variate distribution functions represented through a known copula function, $\mathcal{C} : [0,1]^J \longrightarrow [0,1]$. That is, for any $k \in \mathcal{K} \backslash \{K_1, \ldots, K_J\}$, it holds that

$$F_J(\boldsymbol{\eta}_k) \equiv \mathcal{C}_{\boldsymbol{\gamma}}(F_{1,1}(\eta_{1,k_1}), \ldots, F_{1,J}(\eta_{J,k_J}); \boldsymbol{\gamma}) \tag{4.2}$$

is the cdf with continuous 1-variate marginals $F_{1,j}$, $j \in \mathcal{J}$, and a $K_{\mathcal{C}}$-dimensional association parameter vector $\boldsymbol{\gamma}$ (Sklar, 1959). To allow for the dependence structure implied by the copulae to depend on some covariates $(\boldsymbol{x}_{\gamma_{k_{\mathcal{C}}}})_{k_{\mathcal{C}}=1}^{K_{\mathcal{C}}}$, we also link each association parameter in $\boldsymbol{\gamma}$ to additive predictors $\boldsymbol{\eta}_{\boldsymbol{\gamma}} := (\eta_{\boldsymbol{\gamma},1}, \ldots, \eta_{\boldsymbol{\gamma},K_{\mathcal{C}}})^\top \in \mathbb{R}^{K_{\mathcal{C}}}$ in the spirit of distributional regression (e.g., Klein and Kneib, 2015). In particular, for a given strictly increasing function $h_{k_{\mathcal{C}}}$ that maps the linear predictors into the space of the $k_{\mathcal{C}}$-copula association parameter, we have

$$\gamma_{k_{\mathcal{C}}} = h_{k_{\mathcal{C}}}(\eta_{\gamma,k_{\mathcal{C}}}(\boldsymbol{x}_{\gamma,k_{\mathcal{C}}})) \Longleftrightarrow \gamma^*_{k_{\mathcal{C}}} := h^{-1}_{k_{\mathcal{C}}}(\gamma_{k_{\mathcal{C}}}) = \eta_{\gamma_{k,\mathcal{C}}}(\boldsymbol{x}_{\gamma,k_{\mathcal{C}}}),$$

and $\boldsymbol{\gamma}^* := (\gamma^*_{k_{\mathcal{C}}})_{k_{\mathcal{C}}}$.

We next account for the predictors' specification under a number of covariate effects (e.g., parametric, non-parametric and spatial) before discussing the representation of dichotomous and ordinal polychotomous responses within (4.1). Some models of particular interest in applied research are then discussed in Section 4.3.

4.2.1 Additive Predictors

Let us assume that we observe a sample of size n, $\{\mathrm{y}_{j,i}, \mathbf{x}_{j,i}\}_i$ for every $j \in \mathcal{J}$, whose covariates are collected in the regression matrices $\mathbf{X}_j$ and $\mathbf{X}_{\gamma_{k_{\mathcal{C}}}}$ of dimensions $n \times M_j$ and $n \times M_{\gamma_{k_{\mathcal{C}}}}$. An overall specification of additive predictors is then given by

$$\mathbf{X}_j := (\mathbf{X}^p_{[j]}, \mathbf{X}_{[j,1]}, \ldots, \mathbf{X}_{[j,L_J]}, \mathbf{X}^s_{[j,1]}) \quad \text{and} \quad \bar{\boldsymbol{\beta}}_j := \mathrm{vec}(\boldsymbol{\beta}^p_{[j]}, \boldsymbol{\beta}_{[j,1]}, \ldots, \boldsymbol{\beta}_{[j,L_j]}, \boldsymbol{\beta}^s_{[j,1]}) \in \mathbb{R}^{M_j},$$

$j \in \{1, \ldots, J, \gamma_1, \ldots, \gamma_{k_{\mathcal{C}}}\}$, where the superscripts "p" and "s" refer to parametric and spatial covariate effects, respectively. The proceeding discussion describes the construction of L_j smooth curves representing the effects of the continuous covariate $\mathrm{v}_{j,l_j} \in \mathbb{R}$ on the j-th response, $s_{j,l_j} : \mathbb{R} \longrightarrow \mathbb{R}$, for $l_j = 1, \ldots, L_j$. They are estimated using penalized regression splines in the approach popularised in the literature by Eilers and Marx (1996).

The underlying idea of this method is to approximate each curve by a linear combination of known basis spline functions, b_{j,l_j,h_j}, for $h_j = 1, \ldots, H_j$, and unknown regression parameters to be estimated within the system, δ_{j,l_j,h_j}. In our notation, h_j is employed to count the bases, as delimited by some knot points in the interior of the interval $[\mathrm{v}_{j,l_j,(1)}, \mathrm{v}_{j,l_j,(n)}]$ for every j. Upon defining $\boldsymbol{b}_{j,l_j}(\mathrm{v}_{j,l_j,i}) := (b_{j,l_j,1}(\mathrm{v}_{j,l_j,i}), \ldots, b_{j,l_j,H_j}(\mathrm{v}_{j,l_j,i}))^\top$, and $\boldsymbol{\delta}_{j,l_j}$ the corresponding H_j-dimensional vector of parameters associated with the smooths, the ap-

proximation

$$s_{j,l_j}(\mathrm{v}_{j,l_j,i}) \approx \boldsymbol{\delta}_{j,l_j}^\top \boldsymbol{b}_{j,l_j}(\mathrm{v}_{j,l_j,i}) \in \mathbb{R}, \quad l_j = 1, \ldots, L_j, \quad j \in \mathcal{J}$$

holds. In particular, the evaluation of $\boldsymbol{b}_{j,l_j}(\mathrm{v}_{j,l_j,i})$ for each i yields H_j curves – encompassing different degrees of complexity – that give, once multiplied by some real-valued parameter vector and then summed, an estimated curve for s_{j,l_j}. Basis functions are usually chosen to have convenient mathematical properties and good numerical stability: possible instances include B-splines, cubic regression and low-rank thin plate regression splines (e.g., Ruppert et al., 2003 and Wood, 2003). Since smooth curves are only identifiable up to a constant term, a centering constraint such as $\sum_i s_{j,l_j}(\mathrm{z}_{j,l_j,i}) = 0$ for every l_j has to be imposed. Hence, for non-parametric covariate effects, we have

$$\mathbf{X}_{[j,l_j]} := (\boldsymbol{b}_{j,l_j}(\mathrm{v}_{j,l_j,1}) | \cdots | \boldsymbol{b}_{j,l_j}(\mathrm{v}_{j,l_j,n}))^\top \in \mathbb{R}^{n \times H_j} \quad \text{and} \quad \boldsymbol{\beta}_{[j,l_j]} := \boldsymbol{\delta}_{j,l_j} \in \mathbb{R}^{H_j}.$$

We finally complete the model specification with a ridge-type penalisation acting on the elements of the parameter vector $\boldsymbol{\vartheta}$. Specifically, for each equation j, we construct the quadratic form $\mathcal{P}_j := \boldsymbol{\beta}_j^\top \overline{\mathbf{S}}_{\boldsymbol{\lambda}_j,j} \boldsymbol{\beta}_j$, where

$$\overline{\mathbf{S}}_j := \mathrm{diag}\left(\mathbf{0}, \overline{\mathbf{S}}_{[j,1]}, \ldots, \overline{\mathbf{S}}_{[j,L_j]}, \overline{\mathbf{S}}_{[j]}^s\right)$$

and $\overline{\mathbf{S}}_{[j]}^p = \mathbf{0}$ indicates that no penalisation is attached to the fully parametric model components. Furthermore, we let $\overline{\mathbf{S}}_{\boldsymbol{\lambda}_j,j} := \mathrm{diag}(\boldsymbol{\lambda}_j)\overline{\mathbf{S}}_j$, with $\boldsymbol{\lambda}_j := \mathrm{diag}(\mathbf{0}, \lambda_{[j,1]}, \ldots, \lambda_{[j,L_j]}, \lambda_{[j]}^s) \in [0, +\infty)^{M_j \times M_j}$ be the array collecting the tuning parameters, whose scope is to control for the trade-off between smoothness and fit in the non-parametric estimation of s_{j,l_j}.

The exact definition of the regression matrices and their corresponding penalties for all the covariate specifications allowed by this framework are given in Table 4.1. In conclusion, upon setting $\overline{\mathbf{S}}_{\boldsymbol{\lambda}} := \mathrm{diag}(\overline{\mathbf{S}}_j)_j$, we define an overall penalty $\mathbf{S}_{\boldsymbol{\lambda}}$ as $\overline{\mathbf{S}}_{\boldsymbol{\lambda}}$ padded with zeros such that $\mathcal{P}_{\boldsymbol{\lambda}} := \boldsymbol{\vartheta}^\top \mathbf{S}_{\boldsymbol{\lambda}} \boldsymbol{\vartheta} = \boldsymbol{\beta}^\top \overline{\mathbf{S}}_{\boldsymbol{\lambda}} \boldsymbol{\beta}$.

Covariate Effect	$\mathbf{X}_{[j]}$	$\boldsymbol{\beta}_{[j]}$	$\mathbf{S}_{[j]}$
Parametric	$(\mathbf{x}_{j,i}^\top)_i$	$\bar{\boldsymbol{\beta}}_j$	$\mathbf{0}$
Random Coefficients			$\mathbf{I}$
Non-parametric	$(\boldsymbol{b}_{j,l_j}(\mathrm{v}_{j,l_j,1})\vert\cdots\vert\boldsymbol{b}_{j,l_j}(\mathrm{v}_{j,l_j,n}))^\top$	$\boldsymbol{\delta}_{j,l_j}$	$\int_{V_{j,l_j}} \boldsymbol{b}''_{j,l_j}(\boldsymbol{b}''_{j,l_j})^\top \mathrm{d}\mathrm{v}_{j,l_j}$
Spatial	$(\mathbb{1}_{i=r_j})_{i,r_j} \in \{0,1\}^{n\times R_j}$	$(\boldsymbol{\beta}_{j,r_j})_{r_j}$	$-\mathbb{1}_{r_j\neq s_j}\mathbb{1}_{s_j\in\delta_{r_j}} + \mathbb{1}_{r_j=s_j}N_{r_j}$

Table 4.1: Model specifications for the j-th response corresponding to different covariate effects. Parametric and random coefficient differ only for the penalty matrix: in the latter case it is compatible with coefficient distributed as iid normal with unknown variance (e.g., Wood, 2006). Spatial covariate effects assume R_j discrete adjoint geographical regions indexed by r_j; for any two regions r_j and s_j, δ_{r_j} denotes the set of regions adjacent to r, and $N_r := \#(\delta_r)$. For details please refer to Rue and Held (2005) or Klein et al. (2015).

4.2.2 Discrete Response Representation

Models for discrete outcomes define the linear predictors corresponding to the j-th regression matrix as

$$\begin{aligned}
\boldsymbol{\eta}_j &= \boldsymbol{c}_j - \mathbf{X}^p_{[j]}\boldsymbol{\beta}^p_{[j]} - \cdots - \mathbf{X}^s_{[j]}\boldsymbol{\beta}^s_{[j]} = \boldsymbol{c}_j - \mathbf{X}_j\bar{\boldsymbol{\beta}}_j = \mathbf{Z}_j\boldsymbol{\beta}_j \quad \in \mathbb{R}^n \quad j = 1,2 \\
\boldsymbol{\eta}_{\gamma_{k_\mathcal{C}}} &= \mathbf{X}_{\gamma_{k_\mathcal{C}}}\boldsymbol{\beta}_{\gamma_{k_\mathcal{C}}} \quad \in \mathbb{R}^n \quad \gamma_{k_\mathcal{C}} = \gamma_1,\ldots,\gamma_{K_\mathcal{C}}
\end{aligned}$$

where $\mathbf{Z}_j := (\mathbb{I}_j, -\mathbf{X}_j)$ is the design matrix, $\boldsymbol{\beta}_j := \mathrm{vec}(\boldsymbol{c}_{j,k_j}, \bar{\boldsymbol{\beta}}_j)$, with $\mathbb{I}_j := (\mathbb{1}_{\mathrm{y}_{j,i}=k_j})_{i,k_j} \in \{0,1\}^{n\times K_j-1}$ and $\boldsymbol{c}_{j,k_j} := (c_{j,1},\ldots,c_{j,K_j-1})^\top \in \mathbb{R}^{K_j-1}$ the vector of cut points, $c_{j,K_j} = \infty$. This representation is suitable for both dichotomous and ordinal polychotomous responses, with the caveat that a dichotomous random variable Y_j, defined on $\mathcal{K}_j = \{0,1\}$, usually sets the only cut point to $c_{j,k_j} = 0$. In this case, the parameter vector can be augmented with an intercept term yielding

$$\mathbf{Z}_j = (\mathbf{1}_n, -\mathbf{X}_j) \quad \text{and} \quad \boldsymbol{\beta}_j = \mathrm{vec}(\beta_0, \bar{\boldsymbol{\beta}}_j)$$

for any $\mathrm{y}_{j,i} = k_j$ and $k_j \in \mathcal{K}_j \setminus \{K_j\}$.

Equations (4.1) and (4.2) can now be specialised to account for different response types under the map $\boldsymbol{r}$. To this end, we distinguish three instances representable within the proposed framework.

Case 1: $\mathcal{K} = \{0,1\}^J$**.** For dichotomous responses r is specified such that $r_j : \pi_k \mapsto \pi_k$, the identity map; therefore it holds

$$\boldsymbol{\pi} = \mathcal{C}_{\boldsymbol{\gamma}}(F_{1,1}(\mathbf{Z}_1\boldsymbol{\beta}_1),\ldots,F_{1,J}(\mathbf{Z}_J\boldsymbol{\beta}_J)),$$

where $\boldsymbol{\pi} := (\pi_1, \ldots, \pi_n)^\top \in (0,1)^n$. Multivariate logit/probit models belong to this instance.

Case 2: $\#(\mathcal{K}_j) > 2$ for all $j \in \mathcal{J}$, $(\mathcal{K}_j, \preceq)$ totally ordered. This corresponds to observing ordinal polychotomous responses for any j, for which $r(\pi_{j,k_j}) = \pi_{j,1} + \cdots + \pi_{j,k_j}$ and

$$\sum_{\tilde{k}_1 \preceq k_1} \cdots \sum_{\tilde{k}_J \preceq k_J} \boldsymbol{\pi}(\tilde{k}) = \mathcal{C}_{\boldsymbol{\gamma}}(F_{1,1}(\mathbf{Z}_1\boldsymbol{\beta}_1), \ldots, F_{1,J}(\mathbf{Z}_J\boldsymbol{\beta}_J)), \quad \tilde{k} \in \mathcal{K}. \tag{4.3}$$

The notation $\boldsymbol{\pi}(\tilde{k})$ stresses the fact that the i-th component of $\boldsymbol{\pi}$ depends on the i-th configuration of $\tilde{k}$, namely $\boldsymbol{\pi}(\tilde{k}) := (\pi_i(\tilde{k}^{(i)}))_i$. Be aware that the left-hand side of (4.3) assumes that the elements of $\mathcal{K}$ obey a lexicographical order, that is

$$(\bar{k}_1, \ldots, \bar{k}_J) \preceq (k_1, \ldots, k_J) \Longleftrightarrow (\bar{k}_j \preceq k_j, \forall j) \text{ or } (\bar{k}_j = k_j \ \wedge \ \bar{k}_{\bar{\jmath}} \preceq k_{\bar{\jmath}}, \text{ for some } \bar{\jmath}).$$

Case 3: $\#(\mathcal{K}_j) \geq 2$ for all $j \in \mathcal{J}$. The modelling of mixures of dichotomous and ordinal polychotomous variables is achieved by extending Remark 2 in Chapter 2. Assume without loss of generality that the first $\bar{\jmath} < J$ responses are of the latter type, and the following $J - \bar{\jmath}$ binary. Then we can write

$$(r_J \circ \cdots \circ r_1)(\pi_k) = (r_J \circ \cdots \circ r_{\bar{\jmath}+1}) \circ (r_{\bar{\jmath}} \circ \cdots \circ r_1)(\pi_k) = (r_{\bar{\jmath}} \circ \cdots \circ r_1)(\pi_k),$$

which corresponds to the setting of

$$\sum_{\tilde{k}_1 \preceq k_1} \cdots \sum_{\tilde{k}_{\bar{\jmath}} \preceq k_{\bar{\jmath}}} \boldsymbol{\pi}(\tilde{k}_1, \ldots, \tilde{k}_{\bar{\jmath}}, k_{\bar{\jmath}+1}, \ldots, k_J) = \mathcal{C}_{\boldsymbol{\gamma}}(F_{1,1}(\mathbf{Z}_1\boldsymbol{\beta}_1), \ldots, F_{1,J}(\mathbf{Z}_J\boldsymbol{\beta}_J)).$$

Summary GLMs for discrete data are uniquely characterised by their $(r, F_J, \mathbf{Z})$ form. This can be used to accommodate an enhanced model specification which includes several types of covariate effects, dependence structures between responses, as well as their type.

4.2.3 On the Representation of Ordinal Polychotomous Outcomes

This section aims at investigating in some more details ordinal polychotomous responses within the proposed model specification. Notice first that the given definitions of $r(\pi_{j,k_j})$

pose an immediate constraint on the set $\mathcal{P}$: by assuming $\bar{k}_j \preceq k_j$, we have

$$\begin{aligned} r(\pi_k) &= \sum_{\tilde{k}_1 \leq k_1} \cdots \sum_{\tilde{k}_J \leq k_J} \pi_{\tilde{k}_1,\ldots,\tilde{k}_J} \\ &= \sum_{\tilde{k}_1 \leq k_1} \cdots \sum_{\tilde{k}_j \leq \bar{k}_j} \cdots \sum_{\tilde{k}_J \leq k_J} \pi_{\tilde{k}_1,\ldots,\tilde{k}_J} + \sum_{\tilde{k}_1 \leq k_1} \cdots \sum_{\tilde{k}_j \in [\bar{k}_j, k_j]} \cdots \sum_{\tilde{k}_J \leq k_J} \pi_{\tilde{k}_1,\ldots,\tilde{k}_J} \\ &= r(\pi_{\bar{k}}) + r'(\pi_{\bar{k}}) \geq r(\pi_{\bar{k}}) \end{aligned}$$

since $r'(\pi_k)$ is the sum of probability measures. We also deduce $\bar{k} := (k_1, \ldots, \bar{k}_j, \ldots, k_{J-1}) \preceq (k_1, \ldots, k_j, \ldots, k_J) =: k$ by the assumed lexicographical order. Hence $r(\pi_{\bar{k}}) \leq r(\pi_k)$ for any $\bar{k} \preceq k$, and

$$\mathcal{P} := \{\boldsymbol{r} \in (0,1)^{K-1} | r(\pi_{\bar{k}}) \leq r(\pi_k), \text{ for all } \bar{k} \preceq k \text{ and } \bar{k}, k \in \mathcal{K}\}.$$

If this restriction comes from the very construction of a GLM for ordinal responses (also referred to a Cumulative Link Model, CLM, by McCullagh, 1980), a second one emerges from a general coherency condition. In fact, by inspecting (4.3), we notice that the linear predictors $\{\boldsymbol{\eta}_k\}$ depend on the element k of the discrete ordered set $\mathcal{K}$ that one attempts to model. The sought coherency requires, therefore, the definition of a specific correspondence between the order relations existing in $k \in \mathcal{K}$ with those in $\boldsymbol{\eta}_k \in \mathcal{S}$. This is identified in the order embedding of each $\mathcal{K}_j$ into a relevant subset of the real line as induced by the thresholds $\{c_{j,k_j}\}$. In particular,

Proposition 1. *Let $\varphi : \mathcal{K} \longrightarrow \mathbb{R}^J$ and the finite set $(\mathcal{K}_j, \preceq)$ ordered for every $j \in \mathcal{J}$. Further define φ as*

$$(k_1, \ldots, k_J) \mapsto (\varphi_1(k_1), \ldots, \varphi_J(k_J)) =: (c_1, \ldots, c_J)$$

with $\bar{c}_j \leq c_j$ wherever $\bar{k}_j \preceq k_j$ for every $j \in \mathcal{J}$, and with the elements $k \in \mathcal{K}$ and $\boldsymbol{c}_k \in \mathbb{R}^J$ ordered under a coordinate-wise order. Then φ is an order-embedding.

Proof. See Appendix A.2. ■

The order-embedding allows us to construct non-overlapping hyper-rectangles in $\mathbb{R}^J$ isomorphic to $(k_1, \ldots, k_J) \in \mathcal{K}$. In terms of a multivariate CLM, this is guaranteed by taking the cut points $\{c_{j,k_j}\}$ to be an increasing sequence in k_j for every $j \in \mathcal{J}$, provided that they are the only quantities in the linear predictors depending on the ordered levels of Y_j. To

establish this result, let us consider first

$$\mathcal{Q}_{j,k_j} := \{a \in \mathbb{R} | a \leq \eta_{j,k_j}\} \quad \text{and} \quad \mathcal{R}_{j,k_j} := \Delta_{k_j-1}^{k_j} \mathcal{Q}_{j,k_j} = \{a \in \mathbb{R} | \eta_{j,k_j-1} < a \leq \eta_{j,k_j}\},$$

and construct the set $\mathcal{R} := \mathcal{R}_1 \times \cdots \times \mathcal{R}_J$, with $\mathcal{R}_j := \{\mathcal{R}_{j,1}, \ldots, \mathcal{R}_{j,K_j}\}$; thus it holds:

Proposition 2. *Let φ_j be an order-isomorphism for every $j \in \mathcal{J}$ then the set $\mathcal{R}$ is isomorphic to $\mathcal{K}$.*

Proof. See Appendix A.3. ■

Figure 4.1 shows graphically the idea of this result for $J = 3$. We stress that Proposition 1 and the fact that $\{\boldsymbol{\eta}_k\}$ carries the induced order of $\{\boldsymbol{c}_k\}$ imply the isomorphism $\{\boldsymbol{\eta}_k\} \cong \{k\}$, meaning that there exists a bijection $\phi : \mathcal{K} \longrightarrow \mathcal{S}$ such that $\phi(\bar{k}) := \boldsymbol{\eta}_{\bar{k}} \preceq \boldsymbol{\eta}_k =: \phi(k)$ in $\mathcal{S}$ if and only if $\bar{k} \preceq k$ in $\mathcal{K}$. In this case, the domain of $\mathcal{F}$ is also restricted to

$$\mathcal{S} := \{\boldsymbol{\eta} \in \mathbb{R}^{K-1} | \boldsymbol{\eta}_{\bar{k}} \leq \boldsymbol{\eta}_k, \text{ for all } \bar{k} \preceq k \text{ and } \bar{k}, k \in \mathcal{K}\}.$$

Proposition 2 justifies the interpretation of the discrete $\boldsymbol{Y}$ as a coarse version of a generating continuous latent random vector $\boldsymbol{Y}^* := (Y_1^*, \ldots, Y_J^*)^\top$ in $\mathbb{R}^J$ as originally advocated by McKelvey and Zavoina (1975). Specifically, let $\boldsymbol{\varepsilon} \in \mathbb{R}^J$ be the idiosyncratic components of a regression model of $\boldsymbol{Y}^*$ onto the columns of the design matrices $\{\mathbf{X}_j\}$, then one can write the equivalence

$$\{\boldsymbol{Y} = k\} \Longleftrightarrow \{\boldsymbol{\eta}_{k-1} < \boldsymbol{\varepsilon} \leq \boldsymbol{\eta}_k\},$$

where the right-hand side is intended component-wise as $\eta_{j,k_j-1} < \varepsilon_j \leq \eta_{j,k_j}$ for every $j \in \mathcal{J}$.

In the next section, we qualify the generic framework to describe a class of models widely used in applied research, and we show how it can be represented within the proposed $(r, F_J, \mathbf{Z})$ frame. In this way, these models can be extended beyond a purely parametric functional form of the covariate effects along the lines of the proposals described in this chapter. Moreover, because of the generic approach we take, their estimation and inference will follow as a direct consequence of those of the multivariate penalized GLM for discrete responses.

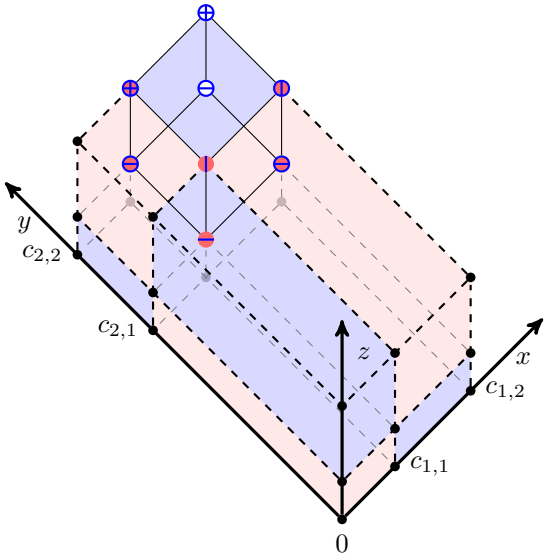

Figure 4.1: A graphical illustration of the construction of $\{R_j\}$ in a subset of $\mathbb{R}^3$. Under φ_j an order-embedding for every $j = 1, 2, 3$, the cut points imply non-overlapping rectangles on $[0, c_{1,2}] \times [0, c_{2,2}] \times [0, c_{3,2}]$. The isomorphism of R and $\mathcal{K}$ (pictured as the lattice top in the figure) is established for any $J < \infty$ in Proposition 2. The c_{j,k_j}'s depicted are the cut points; the ones referring to $j = 3$ correspond to the black dots on the z-axis.

4.3 Some Bivariate Models of Applied Interest

The statistical analysis of observational data may be difficult as they often depart from the ideal conditions underlying any (also rather simple) regression model. They are commonly characterised by a lack of randomisation that may result either in the non-random selection of individuals in the sample, or even in the non-random allocation of a predictor of interest among the population. The former issue is commonly referred to as *non-random sample selection*, and arises whenever individuals select themselves in or out of the relevant sample. It is often the case that some factors that determine the membership to the selected sample are also associated with those that define the outcome itself. In the empirical illustration accompanying this work, and concerning the estimation of the HIV prevalence in Zambia, the refusal of people to be tested for the virus is believed to be induced by variables associated to their HIV status. For example, respondents may already know or correctly predict their seropositivity and so fear that others may learn about that if tested.

The latter instance is regarded instead as a form of *endogeneity*, as it is denominated in the econometric literature. It may stem from different sources, including, but not limited to, direct unmeasured confounding. Wooldridge (2002) discusses in detail several generating sources of endogeneity, thus we refer to him for a more thoughtful illustration of the topic. Unmeasured confounding arises whenever a common background variable affects simultane-

ously both the outcome of interest and one of its regressors, but it is not readily observable or quantifiable by the researcher. The affected covariate is then termed endogenous, and its relationship on the outcome results confounded. A pedagogical example is the estimation of the effect of education on wages. Both variables can be co-determined by factors such as personal ability and motivation that are likely to be explainable by individual's level of education and salary, but hardly measurable (see for example Imbens, 2014 for an interesting survey on the topic).

When not accounted for, non-random sample selection and endogeneity can both lead to inconsistent estimates for all model parameters. To deal with these issues, in some early works Heckman (1978, 1979) devised a two-step estimation procedure for a prototypical recursive bivariate system of equations in a dichotomous responses setting comprising the variables $\boldsymbol{Y} = (Y_1, Y_2)^\top$, with $Y_2(Y_1)$. His proposals specified a binary rule for the observability of the outcome of interest, Y_2, for the non-random sample selection case and, under endogeneity, related the conditional mean of the endogenous regressor, Y_1, to various other predictors. In either scenario, the identification of the true association between the elements of $\boldsymbol{Y}$ would require to be able to qualify the dependence of Y_1 on a relevant variable which is assumed to be independent of both $Y_2|Y_1$ and the unmeasured confounder(s).

Unmeasured Confounding Let $J = 2$ and consider a recursive bivariate structure for the random vector $(Y_1, Y_2(Y_1))^\top \in \mathcal{K}_1 \times \mathcal{K}_2$. The dependence on the first response yields a modification of the linear predictor, which comprises now

$$\mathbf{Z}_2 = (\mathbb{I}_2, -(\mathbf{y}_1, \mathbf{X}_2)) \quad \text{and} \quad \boldsymbol{\beta}_2 = \text{vec}(\boldsymbol{c}_{2,k_2}, \psi, \bar{\boldsymbol{\beta}}_2), \tag{4.4}$$

where $\mathbf{y}_1 := (\text{y}_{1,i})_i$ and $\psi \in \mathbb{R}$ is the coefficient linking Y_1 to Y_2. It is worth stressing that the above structure affects only the definition of $\mathbf{Z}$ in the triplet $(r, F_2, \mathbf{Z})$, and the recursion occurs at the level of manifest Y_1. That is, if one interprets the *intention* towards a manifest discrete outcome (the *observed action*) as the result of an underlying choice mechanism as described by Y_j^*, then (4.4) really describes the effect of an observed endogenous variable (where the intentions have been revealed by the actual choices undertaken) on the discrete response Y_2. This is not the only possibility though. In fact, as Vossmeyer (2014) has recently pointed out, it may also happen that a researcher is genuinely interested in modelling a latent endogenous predictor to be determinant of the intensions about Y_2. Specifically, this corresponds to the instance $Y_2(Y_1^*)$, as previously studied in Chapter 2. A discussion about

the distinctive features of these different modelling strategies can be found in Vossmeyer (2014), who also introduced a formal Bayesian model comparison framework to test these two competing models against the observed data.

Nested models when Y_1 and Y_2 are both dichotomous are discussed in Marra and Radice (2011) and, more recently, in Radice et al. (2015), to which we refer for further details.

Non-random Sample Selection This instance assumes that the outcome $Y_2 \in \mathcal{K}_2$ is observed if and only if $\{0,1\} \ni Y_1 = 1$, whereas it is labelled missing otherwise. As a consequence, the vector $\bar{\boldsymbol{\pi}}$ is also constrained. In fact, every element π_{0,k_2} is now not a sensible quantity for any k_2, since it refers to a missing value in the realisation of Y_2. Hence, one can only model the corresponding marginal probability $\pi_{0\cdot}$, which is translated mathematically into the map $\mathcal{M} \longrightarrow \mathcal{M}^s$, with

$$\mathcal{M}^s := \{\bar{\boldsymbol{\pi}}^s \in (0,1)^{K_2} | \mathbf{1}^\top \bar{\boldsymbol{\pi}}^s < 1\},$$

defined by $\pi_{0,k_2} \mapsto \sum_{\tilde{k}_2 \in \mathcal{K}_2} \pi_{0,\tilde{k}_2} =: \pi_{0\cdot}$, and $\pi_{1,k_2} \mapsto \pi_{1,k_2}$ for any $k_2 \in \mathcal{K}_2 \setminus \{K_2\}$. In complete analogy with the general case, if $\bar{\boldsymbol{\pi}}^s$ is augmented with π_{1,K_2}, the components of the resulting vector will sum up to the unity. Hence, the $(r, F_2, \mathbf{Z})$ representation of this generic sample selection model requires a peculiar definition of the function $\boldsymbol{r}$, as depending on the type of the response Y_2. In particular, we have

$$\begin{array}{rcll} \boldsymbol{r}(\bar{\boldsymbol{\pi}}^s) & = & (\pi_{0\cdot}, \pi_{1,0})^\top & (\#(\mathcal{K}_2) = 2) \\ \boldsymbol{r}(\bar{\boldsymbol{\pi}}^s) & = & (\pi_{0\cdot}, \pi_{1,1}, \ldots, \pi_{1,1} + \cdots + \pi_{1,K_2-1})^\top & (\#(\mathcal{K}_2) > 2) \end{array}.$$

For example, by letting $F_{i,j} = \Phi$ for any j, the Standard Normal cdf, and $\mathcal{C}_\rho$ be the 2-variate Gaussian copula function with correlation coefficient ρ, the recursive versions of the bivariate probit (e.g., Heckman, 1979, Marra and Radice, 2013) and ordered probit models (Miranda and Rabe-Hesketh, 2006) are derived. Notice that in the empirical application of Section 4.5 we consider the dichotomous copula regression model given by

$$(\mathcal{C}_\gamma \circ \mathcal{F})(\boldsymbol{\eta}) = (\mathcal{C}_2(\Phi(-\eta_1), \Phi(\infty); \gamma(\boldsymbol{x}_\gamma)), \mathcal{C}_2(\Phi(\eta_1), \Phi(-\eta_2); \gamma(\boldsymbol{x}_\gamma)))^\top,$$

where $\mathcal{C}_2$ denotes a 2-variate copula function in the Archimedean class, with association parameter γ.

4.4 Elements of Estimation

Let the conditional distribution of $(\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x})$ obey a Categorical distribution with mass function

$$f_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{x}) = \prod_{k\in\mathcal{K}} \pi_k(\boldsymbol{x})^{\mathbb{1}_{\boldsymbol{y}=k}}, \tag{4.5}$$

where $\mathbb{1}_{\boldsymbol{y}=k}$ is a Boolean function that takes value 1 if $(y_1 = k_1 \wedge \cdots \wedge y_J = k_J)$ and 0 otherwise. After having re-defined the response vector $\bar{\boldsymbol{y}} = (\mathbb{1}_{\boldsymbol{y}=1}, \ldots, \mathbb{1}_{\boldsymbol{y}=K})^\top$, (4.5) can be written as

$$f_{\boldsymbol{Y}|\boldsymbol{X}}(\bar{\boldsymbol{y}}|\boldsymbol{x}) = \exp\left\{\bar{\boldsymbol{y}}^\top\boldsymbol{\theta} - b(\boldsymbol{\theta})\right\},$$

with

$$\theta_k = \ln\left\{\frac{\pi_k}{1-\sum_k \pi_k}\right\},\ \theta_K = 0 \ \text{ and } \ b(\boldsymbol{\theta}) = \ln\left\{1+\sum_k \exp\{\theta_k\}\right\}.$$

Hence, the assumed distribution can be expressed in the exponential form and all the standard properties implied by this family follow immediately. Under usual assumptions, equation (4.5) can also be used to derive the log-likelihood function of any multivariate model for discrete data admitting a $(r, F_J, \mathbf{Z})$ representation.

By denoting $\ell_i(\boldsymbol{\vartheta})$ the contribution of the i-th observation to the log-likelihood, the iterative application of the chain rule gives

$$\nabla_{\boldsymbol{\vartheta}}\ell_i(\boldsymbol{\vartheta}) = \frac{\partial\boldsymbol{\eta}_k}{\partial\boldsymbol{\vartheta}}\left(\frac{\partial\mathcal{F}_k}{\partial\boldsymbol{\eta}_k}\frac{\partial\mathcal{C}_{\boldsymbol{\gamma}}}{\partial\mathcal{F}_k}\frac{\partial\pi_k}{\partial\boldsymbol{r}_k}\frac{\partial\theta_k}{\partial\pi_k}\frac{\partial\ell_i}{\partial\theta_k}\right) = \mathbf{D}_i^\top\mathbf{u}_i \qquad \text{and} \qquad \nabla_{\boldsymbol{\vartheta}\boldsymbol{\vartheta}^\top}\ell_i(\boldsymbol{\vartheta}) = \mathbf{D}_i^\top\mathbf{W}_i\mathbf{D}_i + \mathbf{K}_i,$$

with

$$\begin{aligned}\mathbf{W}_i \;=\; & \left[\frac{\partial\boldsymbol{F}_k}{\partial\boldsymbol{\eta}_k}\frac{\partial^2\mathcal{C}_{\boldsymbol{\gamma}}}{\partial\boldsymbol{F}_k\partial\boldsymbol{F}_K^\top}\left(\frac{\partial\boldsymbol{F}_k}{\partial\boldsymbol{\eta}_k}\right)^\top + \frac{\partial^2\boldsymbol{F}_k}{\partial\boldsymbol{\eta}_k\partial\boldsymbol{\eta}_k^\top}\frac{\partial\mathcal{C}_{\boldsymbol{\gamma}}}{\partial\boldsymbol{F}_k}\right]\frac{\partial\pi_k}{\partial\boldsymbol{r}_k}\frac{\partial\theta_k}{\partial\pi_k}\frac{\partial\ell_i}{\partial\theta_k} \\ & + \frac{\partial\boldsymbol{F}_k}{\partial\boldsymbol{\eta}_k}\frac{\partial\mathcal{C}_{\boldsymbol{\gamma}}}{\partial\boldsymbol{F}_k}\frac{\partial\pi_k}{\partial\boldsymbol{r}_k}\left(\frac{\partial\boldsymbol{F}_k}{\partial\boldsymbol{\eta}_k}\frac{\partial\mathcal{C}_{\boldsymbol{\gamma}}}{\partial\boldsymbol{F}_k}\frac{\partial\pi_k}{\partial\boldsymbol{r}_k}\right)^\top\left[\frac{\partial^2\theta_k}{\partial\pi_k^2}\frac{\partial\ell_i}{\partial\theta_k} + \left(\frac{\partial\theta_k}{\partial\pi_k}\right)^2\frac{\partial^2\ell_i}{\partial\theta_k^2}\right].\end{aligned}$$

Notice that the above expressions are the analogous of those derived by Green (1984) in the context of iterative re-weighted least squares (IRLS) estimation of likelihood functions. Indeed, the baseline model is rather similar, with the sole relevant differences being the acknowledgment that only in some special cases $r(\pi_k) = \pi_k$, and the introduction of a copula function in the model specification. In particular, wherever r is the identity map, $\mathbf{u}_i$ reduces to the same simplified expression, $\partial\ell_i/\partial\boldsymbol{\eta}_k$, that appears in Green (1984).

Each individual matrix is finally aggregated into appropriate arrays to get a global repre-

sentation of score and Hessian as follows: $\mathbf{D} := (\mathbf{D}_1^\top | \cdots | \mathbf{D}_n^\top | \mathbf{I}_p)^\top$, $\mathbf{u} := \text{vec}(\mathbf{u}_1, \ldots, \mathbf{u}_n, \mathbf{0}_p)$, $-\mathbf{W} := \text{diag}(\mathbf{W}_1, \ldots, \mathbf{W}_n, \mathbf{K})$, $\mathbf{K} := \sum_i \mathbf{K}_i$, so that $\nabla_{\boldsymbol{\vartheta}} \ell(\boldsymbol{\vartheta}) = \mathbf{D}^\top \mathbf{u}$ and $\nabla_{\boldsymbol{\vartheta}\boldsymbol{\vartheta}^\top} \ell(\boldsymbol{\vartheta}) = -\mathbf{D}^\top \mathbf{W} \mathbf{D}$.

What's next? The setting up of the penalized log-likelihood function allows us to base estimation, smoothing parameters' selection and inference in a similar way as in Chapters 2 and 3. The corresponding procedures are omitted to avoid further repetitions (and preserve my sleep).

4.5 Real Data Illustration: HIV Prevalence in Zambia

We illustrate now the proposed framework through the estimation of a non-random sample selection model. In doing that, we specialise our structure to describe a bivariate probit regression with association parameter explained by an additive linear predictor. This feature is attractive in the context of unmeasured confounding as it allows to account for various degrees of non-random sample selection across observations, and it helps to explain how the association between the relevant outcomes is affected by common unobservables for different individuals and covariates.

The model is applied to data from the 2007 Zambia Demographic Health Survey (DHS) to estimate flexibly the prevalence of HIV in the Zambian male population. Our analysis then complements the study of McGovern et al. (2015) through the inclusion of non-parametric covariate effects, and the specification of the aforementioned elements proper of a distributional regression. The following discussion is further extended by Marra et al. (2015), to which we refer the reader for more extensive and thoughtful argumentations. All the relevant computations presented in the study are performed in the `R` environment (R Development Core Team, 2015) using the package `SemiParBIVProbit` (Marra and Radice, 2015) which implements the ideas discussed in this article for the binary case.

4.5.1 Background and Results

HIV prevalence in a population is defined as the fraction of people who are infected by the virus or, equivalently, as the probability that a randomly drawn individual has the disease. Accurate estimation of the HIV prevalence is essential to policy makers to develop effective control programmes and interventions. Only in recent years, however, in countries where the diffusion of the virus is generalised epidemic, the lack of available administrative data has

been overcome by the intensive use of population-based surveys (Boerma et al., 2003). This is an important new source of data. Prior to their introduction, in fact, national estimates have prevalently relied on a number of sentinel antenatal clinics, whose data usually present different sources of bias (UNAIDS-World Health Organization, 2007). First of all, their samples are based only on sexually active women who are pregnant and attend a clinic; secondly, the location of the facilities, mostly concentrated in urban areas, may also induce a bias even in the subpopulation of pregnant women. These points have been elucidated and discussed with greater details in Montana et al. (2008) and Arpino et al. (2014), among the others.

However, participation rates for HIV testing in national surveys are generally low, and ranges from 72% for men to 77% for women in the 2007 Zambia DHS (Hogan et al., 2012), although even lower peaks are recorded in the 2004 Malawi DHS (63% and 70%, respectively). There are potentially many reasons inducing this pattern, including concerns, lack of incentive to participate, survey fatigue or migration of those targeted for interview (Gersovitz, 2011; Sterck, 2013; McGovern et al., 2015). Missing data on respondents' HIV status represent therefore a not necessarily less severe cause of bias than the ones mentioned above. This case study focuses on refusal to be tested for HIV, which is commonly regarded as the main reason of missingness in surveys.

In this scenario, the use of imputation or weighting techniques is likely to produce biased estimates if the selection mechanism does not occur at random, an assumption that is violated wherever the reasons of the refusal to test are caused by some unobserved factors. This is the case, for example, of individuals who refuse to screen because they already know (or correctly predict) their HIV status, and fear others will learn about their seropositivity if they participate in the survey (McGovern et al., 2015). The framework introduced in this article allows us to estimate a Heckman-type selection model which is able to account for the possibility that data are missing not at random. In particular, this is achieved by modelling item non-response as a function of unobserved variables that also affect the individual HIV status, and specifying the selection mechanism together with an assumption on the distribution of the unobservables. To foster the identification of the causal mechanism in the study, an exclusion restriction is also imposed. We then qualify the dependence of the missing data mechanism on a relevant variable which is assumed independent of both the outcome of interest, given the willingness to take the test, and the unobservables. This factor is usually labelled instrument in econometrics and epidemiology, and interviewer identity is

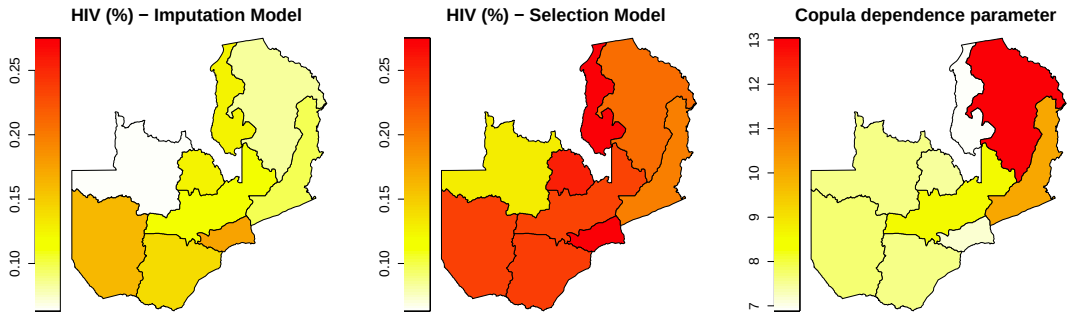

Figure 4.2: First two panels: HIV prevalence for the male population in nine of the ten Provinces of Zambia (Northern, Muchinga, as well as part of Eastern have been merged because of the data availability) applying an imputation model not accounting for the possible presence of values missing not at random, and the corresponding estimates when a bivariate model is fitted instead, respectively. Third panel: the estimated absolute values of the association parameter, with range $(1, \infty)$, in a Joe copula rotated counterclockwise of $90°$. The higher its value, the stronger the estimated association between the two equations; that is, the more relevant the impact of neglecting unobservables in the estimation of the HIV prevalence. The spatial effects are obtained here by specifying appropriately the penalty matrix as described in Section 4.2.1.

regarded as a valid instrument to be employed in our scenario. In fact, previous research, including that of Bärnighausen et al. (2011), Hogan et al. (2012), Janssens et al. (2014) and McGovern et al. (2015), has successfully added such a variable, on the grounds that interviewer identity generally predicts consent to be tested, but it is unlikely to affect the actual HIV status once observed confounders are accounted for.

A pictorial representation of the effects on the estimates of applying a non-random sample selection model is reported in Figure 4.2. By comparison with the first map, the second one shows immediately how the simple imputation of the values under a random missingness assumption may severely underestimate the HIV prevalence in the Zambian provinces. The imputation has been conducted by making predictions from a univariate probit model upon discarding the missing values. The third map depicts instead the ways in which the copula association parameter varies among the different regions of the country, and it has been constructed by exploiting its dependence on the geographical location where the survey took place.

Figure 4.3 shows the smooth function estimates for the treatment and outcome equations, along with their different degrees of non-linearities and associated point-wise confidence intervals, when a Joe_{90} copula model is fitted to the Zambia DHS data. The subscript is used to label the corresponding copula's degrees of rotation. Interestingly, the selection of this copula function conforms with the implied negative association between the two

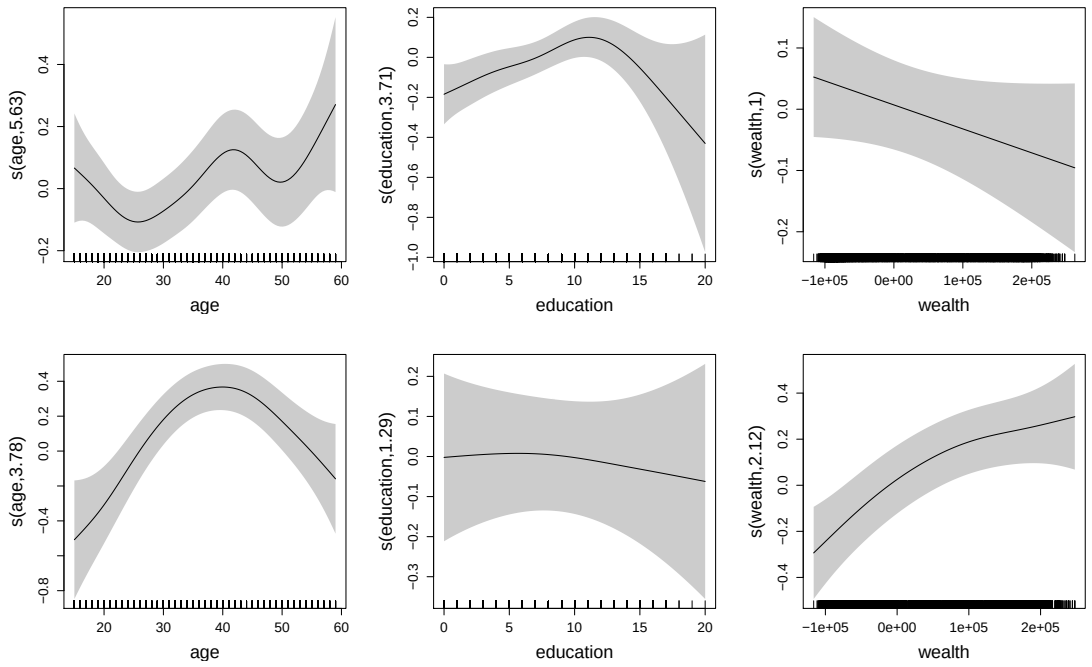

Figure 4.3: Top panel: smooth function estimates and associated 95% point-wise confidence intervals in the treatment equation obtained by applying the Joe_{90} regression spline model on the 2007 Zambia DHS data. Results are plotted on the scale of the linear predictor and the jittered rug plot, at the bottom of each graph, shows the corresponding covariate values. The smooth components are represented using low-rank penalized thin plate regression splines (Wood, 2003) with basis dimensions equal to 10 and penalties based on second order derivatives. The numbers in brackets in the y-axis captions are the effective degrees of freedom of the smooth curves. Bottom panel: estimated smooth functions in the outcome equation.

marginals, as we would expect wherever people refuse to be tested on the basis of some prior knowledge of their HIV+ status. Other existing competitors allowing for the same sign of association include models based on the bivariate Gaussian, Frank, Clayton_{90}, Clayton_{270}, Joe_{270}, Gumbel_{90} and Gumbel_{270} copulae, which are all implemented in `SemiParBIVProbit` and discussed within a system of equations in Radice et al. (2015). As based on information criteria, we found that the Joe_{90} is the best fitting to the male population data, hence our decision to report only selected estimates obtained from this distribution.

As a final remark, we shall stress that the choice of a specific distribution may in principle lead to different estimates of the HIV prevalence (although it seems not to be an issue in our application, e.g. McGovern et al., 2015), and these can be impacted also by the particular functional form of the covariates employed. To deal with this critic, some authors advanced the identification of a *region* (rather than of a singleton) of plausible values in which the parameters of interest necessarily lie, given the available data and the maintained model assumptions. This switch from *point* to *partial* identification is discussed in general terms in

Manski (1995, 2003) and Horowitz and Manski (2000), and applied to a similar HIV context by Arpino et al. (2014). Although theoretically valid and appealing, a major drawback of this approach is the realistic possibility to obtain bounds with large width. That may in turn let the communication of any result to policymakers harsh even in the case where the identifying region is shrunken by the imposition of a monotone instrumental variable.

Acknowledging this issue, our model extends the traditional Heckman-type by accounting for three degrees of flexibility. Namely, the inclusion of non-parametric effects in the representation of the covariate-response functional form, the specification of bivariate copulae to detect more complex dependence structures than what classical distributions usually assume, and the direct modelling of the association parameters in terms of some predictors. It is our hope, in this way, to conjugate both the point and partial identification strengths by providing the researchers with a set of flexible tools aimed at exploring the identifying region widely, and so to make better informed judgments about the robustness of their results wherever a point estimate is sought.

4.6 Discussion

This chapter has devised a generic framework for the representation and estimation of a Generalized Linear Model for a J-dimensional vector of discrete responses, with a ridge-type penalisation term employed in the fitting algorithm. The resulting class of models allowed us to include non-parametric and spatial covariate effects, among others, as represented through the penalty matrix $\mathbf{S}_{\boldsymbol{\lambda}}$. In this way, a baseline multivariate Generalized Additive Model has been effectively extended to encompass different kind of modelling instances within the same unifying framework. In fact, by translating the approach of Peyhardi et al. (2014) to the multivariate case, only the $(r, F_J, \mathbf{Z})$ form and the matrix $\mathbf{S}_{\boldsymbol{\lambda}}$ are formally required to apply the proposed estimation algorithm and related inferential results to different models in the class.

Once the class has been described in some generality, we have introduced a number of bivariate models employed in the literature to account for the possible presence of residual confounding in observational studies. The proposed representation provided us with a flexible machinery able to extend these models in various directions, foremost towards the additive semi-parametric specification of the linear predictors in the spirit of (V)GAMs (Yee and Wild, 1996). This is, *per se*, already a relevant issue in applied research since it permits to

handle a data-driven representation of the covariate-response relationship and so to alleviate a possible source of bias from model mis-specification. Moreover, we have described how the framework can be further specified in order to include multivariate distributions as computed by copulae of univariate marginals.

A further feature illustrated by the chapter has been the direct modelling of any copula association parameter in terms of known predictors. As shown in the analysis of non-random sample selection for the 2007 DHS Zambia dataset. This characteristic is attractive as it allowed us to quantify the strength of the unobservables within the different provinces of the country, and this in turns enabled us to provide new insights about the severity of the non-response issue in the study. In particular, Figure 4.2 showed that the magnitude of the copula association parameter can vary considerably even between geographically close provinces, like Northern and Luapula. On this point, the relevant literature has already stressed that demographic and environmental factors, like the presence of cities or high density housing, may impact the estimates of the HIV prevalence. Hence, the combination of this knowledge with the possibility of letting the association parameters depend on some observed variables seems to us an attractive feature that could be investigated more closely.

As a natural specification of the proposed framework, the practical implementation of models involving ordinal responses are being developed, whereas the estimation of higher dimensional systems of equations is still limited by the necessity of computing multivariate integrals with a good degree of accuracy. In this respect, the exploitation of a more comprehensive class of models for copula distributions may be beneficial, possibly by allowing the non-parametric estimation of the marginals and/or the corresponding copulae. These are only some of the possible avenues of future research that will be undertaken.

Chapter 5

The Ultima Thule

In this thesis I have discussed flexible simultaneous equation regression techniques to account for the presence of unobservables in cross-sectional studies. Methodologies were developed for bivariate systems of ordinal polychotomous response equations with an enhanced functional form representation of the covariate effects. Two new models have been consequently introduced and their computational routines made available to the interested users.

After a brief review of the main concepts of regression splines modelling, Chapter 2 tackled the problem of unobserved confounding in a discrete data setting. A semi-parametric IV estimation procedure has been developed and tested against a number simulated scenarios. They confirmed the validity of the proposed procedures to obtain consistent curve and parameter estimates whenever unobservables are suspected to be present.

In Chapter 3, I described instead the use of a copula semi-parametric regression for the analysis of injury severities in vehicle accidents. By acknowledging the role of common unobservables in multi-party crashes, I studied a flexible SURE model and discussed the benefits of a simultaneous estimation approach compared to various alternatives available from the literature. Although I only considered the instance of Standard Normal marginals, other univariate distributions are in principle allowed and implementable.

The common structure of the above models was finally exploited in Chapter 4, which introduced a generic multivariate representation for discrete responses in penalized GLMs. I showed that some existing flexible models for unmeasured confounding and non-random sample selection are nested in this framework, and a number of other expressible to accommodate enhanced covariate-response relationships. The chapter was intended to form the theoretical basis to specify bivariate models for several types of responses, including nominal, sequential, or potentially any mixture of them.

In the concluding remarks to each chapter, I have illustrated specific possible routes for future research generated by the present thesis. Alongside them, I reckon that the development of a unique `R` package for both `SemiParCLM` and `CopulaCLM` could be a valuable task to be undertaken in the immediate future. In fact, this may ease the diffusion of the computational routines needed to fit these models and, possibly, encourage their use in

applied contexts other than those discussed in the thesis. This would require the definition of a common set of options available to both the models, also considering that Chapter 2 does not currently allow for a copula specification. On this point, the use of a different class of copulae might permit to overcome the limitation of having the distribution of Y_2^* expressed as the convolution of ε_1 and ε_2.

On a different side, the extension of the proposed methodologies to a longitudinal data setting could represent a promising avenue of future research. Moreover, acknowledging that many data are nowadays collected in the form of $p \gg n$, a way to perform variable selection could be of some practical relevance. A possible approach, for example, could be to implement a composition of L_1 and L_2 penalties in the system with an acceptable degree of computational speed. The models could also be extended to incorporate estimation procedures robust to the presence of outliers. On this point, I have the impression that the working linear models employed in the thesis can be appropriately modified in a least trimmed squares regression fashion to achieve robust estimation (Rousseeuw, 1984), although the methods used to estimate the smoothing parameters may not be applied directly. These paths will be investigated during my stay at the European Central Bank.

Now the sunset is approaching Strasbourg and the caffeine on my desk over: finally, quite a good time to state a laical

NUNC DIMITTIS.

Appendix A

Further Technical Results

A.1 Proof of Result (2.12)

In addition to the previous assumptions (i)-(iv), we further assume the following: (v) for every $\vartheta^s \in \boldsymbol{\vartheta}$, $\partial^3/\partial\vartheta^{s3}(\ell_n(\boldsymbol{\vartheta}))$ exists and satisfies for every point $x \in \mathbb{R}$ and every parameter in the neighbourhood of ϑ_0^s: $|\partial^3/\partial\vartheta^{s3}(\ell_n(\boldsymbol{\vartheta}))| \leq M(x)$, with $\mathbb{E}[M(x)|\vartheta_0^s] < \infty$; and let $0 \leq \mathcal{I}(\vartheta_0^s) < \infty$.

Proof. We first set the notation. Let us denote by ϑ^j the j-th component of the parameter vector $\boldsymbol{\vartheta} = (\vartheta^1, \ldots, \vartheta^p)^\top$, and define $\ell_{\mathrm{p},j} := \partial\ell_\mathrm{p}/\partial\vartheta^j$ as the partial derivative of the penalized log-likelihood with respect to ϑ^j; higher order derivatives are denoted subsequently. Also, the "hat" notation $\widehat{\ell}_\mathrm{p}$ stands for $\ell_\mathrm{p}(\widehat{\boldsymbol{\vartheta}})$, while the convention of omitting the listing of parameters is used wherever the relevant quantities are evaluated at the best coefficient $\boldsymbol{\vartheta}_0$, that is $\ell_\mathrm{p} := \ell_\mathrm{p}(\boldsymbol{\vartheta}_0)$.

Using the Einstein summation convention, we expand $\widehat{\ell}_{\mathrm{p},r}$ around $\ell_{\mathrm{p},r}$ using a second order Taylor approximation:

$$0 = \widehat{\ell}_{\mathrm{p},r} = \ell_{\mathrm{p},r} + \ell_{\mathrm{p},rs}(\widehat{\vartheta} - \vartheta_0)^s + \frac{1}{2}\ell_{\mathrm{p},rst}(\widehat{\vartheta} - \vartheta_0)^{st} + \cdots$$

with $(\widehat{\vartheta} - \vartheta_0)^s := \widehat{\vartheta}^s - \vartheta_0^s$ and $(\widehat{\vartheta} - \vartheta_0)^{st} = (\widehat{\vartheta} - \vartheta_0)^s(\widehat{\vartheta} - \vartheta_0)^t$. Solving the above equation for $\widehat{\vartheta} - \vartheta_0$, and denoting by superscripts the inverses of the respective quantities, we get (Barndorff-Nielsen and Cox, 1994):

$$(\widehat{\vartheta} - \vartheta_0)^r = -\ell_\mathrm{p}^{rs}\ell_{\mathrm{p},s} - \frac{1}{2}\ell_\mathrm{p}^{rtv}\ell_{\mathrm{p},u}\ell_{\mathrm{p},w} + \cdots \tag{A.1}$$

where $\ell_\mathrm{p}^{rtv} := \ell_\mathrm{p}^{rs}\ell_\mathrm{p}^{tu}\ell_\mathrm{p}^{vw}\ell_{\mathrm{p},stv}$, and ℓ_p^{rs} is the (r,s)-th element of the inverse observed (penalized) Fisher Information. Equation (A.1) can be simplified as follows (see, for example, Kauermann, 2005): $\ell_{\mathrm{p},rs} := f_{rs}(\lambda) + r_{rs}$, where $f_{rs}(\lambda) := f_{rs}(0) - s_\lambda^{rs}$ is the penalized expected Fisher Information contribution: $f_{rs}(0) := \mathbb{E}[\partial\ell/\partial\vartheta^r\partial\vartheta^s]$, and $r_{rs} := \ell_{rs} - f_{rs}(0)$.

Under assumptions (ii) and (iv) we find that $f_{rs}(\lambda)$ is of asymptotic order $O(n)$, and

that $r_{rs} = O_p(n^{1/2})$ directly from (iii). We can then simplify the first term of (A.1) as

$$\begin{aligned} -\ell_{\mathrm{p}}^{rs} &= \mathbb{E}[\ell_{\mathrm{p},r}\ell_{\mathrm{p},s}]^{-1} + \mathbb{E}[\ell_{\mathrm{p},r}\ell_{\mathrm{p},t}]^{-1}\mathbb{E}[\ell_{\mathrm{p},s}\ell_{\mathrm{p},u}]^{-1}(\mathbb{E}[\ell_{\mathrm{p},t}\ell_{\mathrm{p},u}] + \ell_{p,tu}) \\ &= -f^{rs}(\lambda) + f^{rt}(\lambda)f^{su}(\lambda)(-f_{rs}(\lambda) + \ell_{p,tu}), \end{aligned}$$

that is $\ell_{\mathrm{p}}^{rs} = f^{rs}(\lambda) - f^{rt}(\lambda)f^{su}r_{tu}$; following now the argument of Kauermann et al. (2009) we have

$$\ell_{\mathrm{p}}^{rs} = f^{rs}(\lambda)[1 + O(n^{-1})O_p(n^{1/2})] = f^{rs}(\lambda)[1 + O_p(n^{-1/2})].$$

We next need to characterise the order of ℓ_{p}^{rtv}, which in turn depends on the one of $\ell_{\mathrm{p},stv}$. First note that $\ell_{\mathrm{p},stv} = \ell_{stv}$ from the very construction of the penalized likelihood estimator, so that we can safely apply (v), implying that we can bound in probability the third derivative of the log-likelihood. Then, by the strong law of large numbers, we have that, for almost every sequence of $\{x_1, \ldots, x_n\}$ and every $\boldsymbol{\vartheta} \in \Theta$,

$$|n^{-1}\ell_{stv}| \leq n^{-1}\sum_i M(x_i) \xrightarrow{as} \mathbb{E}[M(x)]$$

as $n \to \infty$, hence $n^{-1}\ell_{stv} = O_p(1)$. It is then implied $\ell_{stv} = O_p(n)$ and, after some tedious computations, $\ell_{\mathrm{p}}^{rtv} = f^{rs}(\lambda)f^{tu}(\lambda)f^{vw}(\lambda)O_p(n) = O_p(n^{-2})$ so that $\ell_{\mathrm{p}}^{rtv}\ell_{\mathrm{p},u}\ell_{\mathrm{p},w} = O_p(n^{-1})$ since $\ell_{\mathrm{p},u} = O_p(n^{1/2}) - o(n^{1/2})$. We also find that $\ell_{\mathrm{p}}^{rs}\ell_{\mathrm{p},s}$ has order $O_p(n^{-1/2}) + o(n^{-1/2})$, that is the second addendum in (A.1) becomes asymptotically negligible compared to $\ell_{\mathrm{p}}^{rs}\ell_{\mathrm{p},s}$. We can then write $(\widehat{\vartheta} - \vartheta_0)^r = -f^{rs}(\lambda)\ell_{\mathrm{p},s}[1 + o_p(1)]$, whose leading terms, in matrix notation, are $\mathbf{F}^{-1}(\boldsymbol{\lambda})(\nabla_{\boldsymbol{\vartheta}_0}\ell(\boldsymbol{\vartheta}_0) - \mathbf{S}_{\boldsymbol{\lambda}}\boldsymbol{\vartheta}_0)$, from which the assertion follows.

The stochastic order of the above terms then stems from $f^{rs}(\lambda)\ell_{\mathrm{p},s} = O_p(n^{-1/2}) + o_p(n^{-1/2}) = O_p(n^{-1/2})$. ■

A.2 Proof of Proposition 1

Fix first $J = 1$ and let $\varphi_j : \mathcal{K}_j \longrightarrow \mathbb{R}$ be such that $\bar{k}_j \mapsto c_{j,\bar{k}_j}$ and $k_j \mapsto c_{j,k_j}$ for any $\bar{k}_j, k_j \in \mathcal{K}_j$, and $c_{j,\bar{k}_j} \leq c_{j,k_j}$ for any $c_{j,\bar{k}_j}, c_{j,k_j} \in \mathbb{R}$. Then φ_j is an order-embedding by construction. Conversely, being φ_j an order-embedding, the image of $\mathcal{K}_j$ under φ_j, $\varphi_j(\mathcal{K}_j) := \{\varphi(k_j) | k_j \in \mathcal{K}_j\}$, is isomorphic to $\mathcal{K}_j$. Therefore, using the definition of φ_j, we also have

$$\{\bar{k}_j \preceq k_j\} \Longleftrightarrow \{c_{j,\bar{k}_j} \leq c_{j,k_j}\}.$$

We establish the argument now for $J > 1$. By definition of order-embedding, we need to show that the map φ is such that $\bar{k} \preceq k$ if and only if $\varphi(\bar{k}) \leq \varphi(k)$ in $\mathbb{R}^J$. Under the coordinate-wise order we also have:

$$\begin{aligned} k := (\bar{k}_1, \ldots, \bar{k}_J) \preceq (k_1, \ldots, k_J) =: \bar{k}_j \quad &\iff \quad \{\bar{k}_j \preceq k\} \iff \{c_{j,\bar{k}_j} \leq c_{j,k_j}\} \\ &\iff \quad \boldsymbol{c}_{\bar{k}} := (c_{1,\bar{k}_1}, \ldots, c_{J,\bar{k}_J}) \leq (c_{1,k_1}, \ldots, c_{J,k_J}) =: \boldsymbol{c}_k \end{aligned}$$

for all $j \in \mathcal{J}$ because we have showed that the condition $c_{j,\bar{k}_j} \leq c_{j,k_j}$ is a criterion for the embedding of the order of $\mathcal{K}_j$ into $\mathbb{R}$, and the assumption of the coordinate-wise ordering. This establishes that $\bar{k} \preceq k$ in $\mathcal{K} \iff \varphi(k) \leq \varphi(\bar{k})$ in $\mathbb{R}^J$, meaning that φ is an order-embedding.

A.3 Proof of Proposition 2

As a preliminary results we have:

Lemma 1. *The set $(\mathcal{K}_j, \preceq)$ is an interval order under the mapping φ_j.*

Proof. The intervals on $\mathbb{R}$ ordered by left-to-right precedence, $\preceq^{\ell r}$, and implied by φ_j are $\{\mathcal{R}_{j,k_j}\}$. The claim then requires the isomorphism between $\mathcal{K}_j$ and $\mathcal{R}_j$. First note that, if $\mathcal{Q}_j \cong \mathcal{K}_j$, it suffices to show that $R_j \cong \mathcal{Q}_j$, where $\mathcal{Q}_j := \{\mathcal{Q}_{j,1}, \ldots, \mathcal{Q}_{j,K_j}\}$. This is, however, an easy statement. Define the function $\phi_j : \mathcal{Q}_j \longrightarrow \mathcal{R}_j$ as $\mathcal{Q}_{j,k_j} \mapsto \mathcal{R}_{j,k_j}$ where the index k_j denotes the same position of the elements into the corresponding sets under a lexicographic order. For any $\mathcal{Q}_{j,\bar{k}_j} \subseteq \mathcal{Q}_{j,k_j}$ we have $\phi_j(\mathcal{Q}_{j\bar{k}_j}) = \mathcal{R}_{j,\bar{k}_j}$ and $\phi_j(\mathcal{Q}_{j,k_j}) = \mathcal{R}_{j,k_j}$, with $\mathcal{R}_{j,\bar{k}_j} \preceq^{\ell r} \mathcal{R}_{j,k_j}$ by the very construction of the interval order relations. The other direction is attained similarly by construction. Hence

$$\mathcal{Q}_{j,\bar{k}_j} \subseteq \mathcal{Q}_{j,k_j} \text{ in } \mathcal{Q}_j \text{ if and only if } \phi_j(\mathcal{Q}_{j,\bar{k}_j}) \preceq^{\ell r} \phi_j(\mathcal{Q}_{j,k_j}) \text{ in } \mathcal{R}_j,$$

so that $\mathcal{Q}_j$ and $\mathcal{R}_j$ are isomorphic. Thus it is left to prove the requirement $\mathcal{Q}_j \cong \mathcal{K}_j$.

To this end, we have already proved in Proposition 1 that $\{k_j\} \cong \{c_{j,k_j}\}$ for any j whereas, for any two $c_{j,\bar{k}_j}, c_{j,k_j} \in \mathbb{R}$, it holds

$$c_{j,\bar{k}_j} \leq c_{j,k_j} \iff \mathcal{Q}_{j,\bar{k}_j} \subseteq \mathcal{Q}_{j,k_j}.$$

Hence $\{k_j\} \cong \{\mathcal{Q}_{j,k_j}\}$ and $\mathcal{K}_j \cong \mathcal{R}_j$. In other words, the map φ_j defined as in Proposition 1

induces non-overlapping intervals on the real line (delimited by monotonic cut points) which are isomorphic to the order relations in $\mathcal{K}_j$. ■

The proof of Proposition 2 then reads as:

Proof of Proposition 2. We first claim that for any $j, \tilde{\jmath} \in \mathcal{J}$ it holds that $\mathcal{K}_j \times \mathcal{K}_{\tilde{\jmath}} \cong \mathcal{R}_j \times \mathcal{R}_{\tilde{\jmath}}$. Then, since both $\mathcal{K}_j \times \mathcal{K}_{\tilde{\jmath}}$ and $\mathcal{R}_j \times \mathcal{R}_{\tilde{\jmath}}$ are ordered, the repeated application of the statement gives the result.

We devote now to the proof of the claim. From Lemma 1 we know that $\mathcal{R}_j \cong \mathcal{K}_j$ for any $j \in \mathcal{J}$ so there exists a function $\varphi_j^{\ell r} : \mathcal{K}_j \longrightarrow \mathcal{R}_j$ such that $k_{\tilde{\jmath}} \preceq k_j$ in $\mathcal{K}_j$ if and only if $\varphi_j^{\ell r}(k_{\tilde{\jmath}}) \preceq^{\ell r} \varphi_j^{\ell r}(k_j)$ in $\mathcal{R}_j$. Since $\mathcal{K}_j$ and $\mathcal{R}_j$ are finite ordered set and $\varphi_j^{\ell r}$ an order-isomorphism, then the latter is equivalent to the condition (Lemma 1.17 in Davey and Priestley, 2002)

$$k_{\tilde{\jmath}} \prec k_j \text{ if and only if } \varphi_j^*(k_{\tilde{\jmath}}) \prec \varphi_j^{\ell r}(k_j),$$

where the symbol $\prec$ denotes the covering relation between elements of an ordered set. Specifically, $k_{\tilde{\jmath}} \prec k_j$ means $k_{\tilde{\jmath}} \preceq k_{\tilde{\jmath}} \prec k_j \Longrightarrow k_{\bar{\jmath}} = k_{\tilde{\jmath}}$. If we denote by R_{j,k_j} the generic element of $\mathcal{R}_j$, we have (Davey and Priestley, 2002, ex. 1.7):

$$\begin{aligned}
& (\mathcal{R}_{j,\bar{k}_j}, \mathcal{R}_{\tilde{\jmath},\bar{k}_{\tilde{\jmath}}}) \prec (\mathcal{R}_{j,k_j}, \mathcal{R}_{\tilde{\jmath},k_{\tilde{\jmath}}}) \\
\iff & (\mathcal{R}_{j,\bar{k}_j} = \mathcal{R}_{j,k_j} \wedge \mathcal{R}_{\tilde{\jmath},\bar{k}_{\tilde{\jmath}}} \prec \mathcal{R}_{\tilde{\jmath},k_{\tilde{\jmath}}}) \ \vee \ (\mathcal{R}_{j,\bar{k}_j} \prec \mathcal{R}_{j,k_j} \wedge \mathcal{R}_{\tilde{\jmath},\bar{k}_{\tilde{\jmath}}} = \mathcal{R}_{\tilde{\jmath},k_{\tilde{\jmath}}}) \\
\iff & (\varphi_j^{\ell r}(\bar{k}_j) = \varphi_j^{\ell r}(k_j) \wedge \varphi_{\tilde{\jmath}}^{\ell r}(\bar{k}_{\tilde{\jmath}}) \prec \varphi_{\tilde{\jmath}}^{\ell r}(k_{\tilde{\jmath}})) \ \vee \ (\varphi_j^{\ell r}(\bar{k}_j) \prec \varphi_j^{\ell r}(k_j) \wedge \varphi_{\tilde{\jmath}}^{\ell r}(\bar{k}_{\tilde{\jmath}}) = \varphi_{\tilde{\jmath}}^{\ell r}(k_{\tilde{\jmath}})) \\
\iff & (\bar{k}_j = k_j \wedge \bar{k}_{\tilde{\jmath}} \prec k_{\tilde{\jmath}}) \ \vee \ (\bar{k}_j \prec k_j \wedge \bar{k}_{\tilde{\jmath}} = k_{\tilde{\jmath}}) \\
\iff & (\bar{k}_j, \bar{k}_{\tilde{\jmath}}) \prec (k_j, k_{\tilde{\jmath}}),
\end{aligned}$$

where the third line stems from the isomorphism of $\mathcal{K}_j$ and $\mathcal{R}_j$ for any $j \in \mathcal{J}$, and the last from the assumption of $\mathcal{K}_j$ ordered. Hence we have

$$(\mathcal{R}_{j,\bar{k}_j}, \mathcal{R}_{\tilde{\jmath},\bar{k}_{\tilde{\jmath}}}) \prec (\mathcal{R}_{j,k_j}, \mathcal{R}_{\tilde{\jmath},k_{\tilde{\jmath}}}) \iff (\bar{k}_j, \bar{k}_{\tilde{\jmath}}) \prec (k_j, k_{\tilde{\jmath}}).$$

Define now the map $\varphi^{\ell r} : \mathcal{K}_j \times \mathcal{K}_{\tilde{\jmath}} \longrightarrow \mathcal{R}_j \times \mathcal{R}_{\tilde{\jmath}}$ as $(k_j, k_{\tilde{\jmath}}) \mapsto (\mathcal{R}_{j,k_j}, \mathcal{R}_{\tilde{\jmath},k_{\tilde{\jmath}}})$, and set $\varphi^{\ell r}(k_j, k_{\tilde{\jmath}}) := (\varphi_j^{\ell r}(k_j), \varphi_{\tilde{\jmath}}^{\ell r}(k_{\tilde{\jmath}}))$ with $\varphi_j^{\ell r}$ and $\varphi_{\tilde{\jmath}}^{\ell r}$ defined as above. Then the statement

$$\varphi^{\ell r}(\bar{k}_j, \bar{k}_{\tilde{\jmath}}) \prec \varphi^{\ell r}(k_j, k_{\tilde{\jmath}}) \text{ in } \mathcal{R}_j \times \mathcal{R}_{\tilde{\jmath}} \text{ if and only if } (\bar{k}_j, \bar{k}_{\tilde{\jmath}}) \prec (k_j, k_{\tilde{\jmath}}) \text{ in } \mathcal{K}_j \times \mathcal{K}_{\tilde{\jmath}},$$

which is equivalent, by the above-mentioned Lemma 1.17, to:

$$\varphi^{\ell r}(\bar{k}_j, \bar{k}_{\tilde{\jmath}}) \preceq^{\ell r} \varphi^{\ell r}(k_j, k_{\tilde{\jmath}}) \text{ in } \mathcal{R}_j \times \mathcal{R}_{\tilde{\jmath}} \text{ if and only if } (\bar{k}_j, \bar{k}_{\tilde{\jmath}}) \preceq (k_j, k_{\tilde{\jmath}}) \text{ in } \mathcal{K}_j \times \mathcal{K}_{\tilde{\jmath}},$$

meaning that $\varphi^{\ell r}$ is an order-isomorphism, and so $\mathcal{R}_j \times \mathcal{R}_{\tilde{\jmath}} \cong \mathcal{K}_j \times \mathcal{K}_{\tilde{\jmath}}$ $\forall j, \tilde{\jmath} \in \mathcal{J}$ as claimed. Therefore it also holds that $\mathcal{R}$ is isomorphic to $\mathcal{K}$. ■

Appendix B

Complementary Materials

B.1 Complements to Chapter 2

B.1.1 Construction of the score and Hessian matrix

As illustrated in the chapter, the score vector is obtained from the matrix multiplication $\mathbf{D}^\top \mathbf{u}$, where

$$\mathbf{D}^\top = \begin{bmatrix} \mathbf{D}_{1,1}^\top & \mathbf{D}_{1,2}^\top & \mathbf{0}^\top & \mathbf{0}^\top & \mathbf{0}^\top \\ \mathbf{D}_{2,1}^\top & \mathbf{D}_{2,2}^\top & \mathbf{0}^\top & \mathbf{0}^\top & \mathbf{0}^\top \\ \mathbf{0}^\top & \mathbf{0}^\top & \mathbf{D}_{3,3}^\top & \mathbf{D}_{3,4}^\top & \mathbf{0}^\top \\ \mathbf{0}^\top & \mathbf{0}^\top & \mathbf{D}_{4,3}^\top & \mathbf{D}_{4,4}^\top & \mathbf{0}^\top \\ \mathbf{D}_{5,1}^\top & \mathbf{D}_{5,1}^\top & \mathbf{D}_{5,3}^\top & \mathbf{D}_{5,3}^\top & \mathbf{0}^\top \\ \mathbf{0}^\top & \mathbf{0}^\top & \mathbf{D}_{6,3}^\top & \mathbf{D}_{6,3}^\top & \mathbf{0}^\top \\ \mathbf{0}^\top & \mathbf{0}^\top & \mathbf{D}_{7,3}^\top & \mathbf{D}_{7,4}^\top & \mathbf{D}_{7,5}^\top \\ \mathbf{0}^\top & \mathbf{0}^\top & \mathbf{D}_{8,3}^\top & \mathbf{D}_{8,4}^\top & \mathbf{D}_{8,5}^\top \end{bmatrix} \quad \text{and} \quad \mathbf{u} = \widetilde{\mathbf{P}} \begin{bmatrix} \mathbf{u}_{1,1} \\ \mathbf{u}_{1,2} \\ \mathbf{u}_{2,1} \\ \mathbf{u}_{2,2} \\ \mathbf{u}_3 \end{bmatrix},$$

with $\widetilde{\mathbf{P}}$ collecting into an appropriate way the term $1/\pi_k$ in (2.6). In particular, each component of the above arrays is given by

- $\vartheta_1 = \tilde{c}_{1,1}$:

$$\begin{aligned} \mathbf{D}_{1,1} &:= \tfrac{\mathrm{d}\boldsymbol{\eta}_{1,k_1}}{\mathrm{d}\tilde{c}_{1,1}} = \mathbb{1}_{k_1 \geq 1}\mathbf{1}_n \qquad k_1 = 1, \ldots, K_1 - 1 \\ \mathbf{D}_{1,2} &:= \tfrac{\mathrm{d}\boldsymbol{\eta}_{1,k_1-1}}{\mathrm{d}\tilde{c}_{1,1}} = \mathbb{1}_{k_1 \geq 2}\mathbf{1}_n \quad k_1 = 2, \ldots, K_1 \end{aligned}$$

- $\boldsymbol{\vartheta}_2^\top = \tilde{\boldsymbol{c}}_{1,h}^\top$, $h = 2, \ldots, K_1 - 1$:

$$\begin{aligned} \mathbf{D}_{2,1} &:= \tfrac{\mathrm{d}\boldsymbol{\eta}_{1,k_1}}{\mathrm{d}\tilde{\boldsymbol{c}}_{1,h}^\top} = 2[\mathbb{1}_{k_1 \geq h}\mathbf{1}_n]\tilde{\boldsymbol{c}}_{1,h}^\top \qquad k_1 = 1, \ldots, K_1 - 1 \\ \mathbf{D}_{2,2} &:= \tfrac{\mathrm{d}\boldsymbol{\eta}_{1,k_1-1}}{\mathrm{d}\tilde{\boldsymbol{c}}_{1,h}^\top} = 2[\mathbb{1}_{k_1 \geq h+1}\mathbf{1}_n]\tilde{\boldsymbol{c}}_{1,h}^\top \quad k_1 = 2, \ldots, K_1 \end{aligned}$$

- $\vartheta_3 = \tilde{c}_{2,1}$:

$$\begin{aligned}
\mathbf{D}_{3,3} &:= \frac{\mathrm{d}\boldsymbol{\eta}_{2,k_2}}{\mathrm{d}\tilde{c}_{2,1}} = \frac{\mathbb{1}_{k_2\geq 1}}{\sqrt{1+2\psi\rho+\psi^2}}\,\mathbf{1}_n \qquad k_2 = 1,\ldots,K_2-1\\
\mathbf{D}_{3,4} &:= \frac{\mathrm{d}\boldsymbol{\eta}_{2,k_2-1}}{\mathrm{d}\tilde{c}_{2,1}} = \frac{\mathbb{1}_{k_2\geq 2}}{\sqrt{1+2\psi\rho+\psi^2}}\,\mathbf{1}_n \quad k_2 = 2,\ldots,K_2
\end{aligned}$$

- $\boldsymbol{\vartheta}_4^\top = \tilde{\boldsymbol{c}}_{2,h}^\top$, $h = 2,\ldots,K_2-1$:

$$\begin{aligned}
\mathbf{D}_{4,3} &:= \frac{\mathrm{d}\boldsymbol{\eta}_{2,k_2}}{\mathrm{d}\tilde{\boldsymbol{c}}_{2,h}^\top} = 2\,\frac{\mathbb{1}_{k_2\geq h}}{\sqrt{1+2\psi\rho+\psi^2}}\,\tilde{\boldsymbol{c}}_{2,h}^\top \qquad k_2 = 1,\ldots,K_2-1\\
\mathbf{D}_{4,4} &:= \frac{\mathrm{d}\boldsymbol{\eta}_{2,k_2-1}}{\mathrm{d}\tilde{\boldsymbol{c}}_{2,h}^\top} = 2\,\frac{\mathbb{1}_{k_2\geq h+1}}{\sqrt{1+2\psi\rho+\psi^2}}\,\tilde{\boldsymbol{c}}_{2,h}^\top \quad k_2 = 2,\ldots,K_2
\end{aligned}$$

- $\boldsymbol{\vartheta}_5^\top = \boldsymbol{\beta}_1^\top$:

$$\begin{aligned}
\mathbf{D}_{5,1} &:= \frac{\mathrm{d}\boldsymbol{\eta}_{1,k_1}}{\mathrm{d}\boldsymbol{\beta}_1^\top} = \frac{\mathrm{d}\boldsymbol{\eta}_{1,k_1-1}}{\mathrm{d}\boldsymbol{\beta}_1^\top} = -\mathbf{X}_1\\
\mathbf{D}_{5,3} &:= \frac{\mathrm{d}\boldsymbol{\eta}_{2,k_2}}{\mathrm{d}\boldsymbol{\beta}_1^\top} = \frac{\mathrm{d}\boldsymbol{\eta}_{2,k_2-1}}{\mathrm{d}\boldsymbol{\beta}_1^\top} = -\frac{\psi}{\sqrt{1+2\psi\rho+\psi^2}}\,\mathbf{X}_1
\end{aligned}$$

- $\boldsymbol{\vartheta}_6^\top = \boldsymbol{\beta}_2^\top$:

$$\mathbf{D}_{6,3} := \frac{\mathrm{d}\boldsymbol{\eta}_{2,k_2}}{\mathrm{d}\boldsymbol{\beta}_2^\top} = \frac{\mathrm{d}\boldsymbol{\eta}_{2,k_2-1}}{\mathrm{d}\boldsymbol{\beta}_2^\top} = -\frac{1}{\sqrt{1+2\psi\rho+\psi^2}}\,\mathbf{X}_2$$

- $\vartheta_7 = \psi$:

$$\begin{aligned}
\mathbf{D}_{7,3} &:= \frac{\mathrm{d}\boldsymbol{\eta}_{2,k_2}}{\mathrm{d}\psi} = -\frac{1}{\sqrt{1+2\psi\rho+\psi^2}}\,(\mathbf{X}_1\boldsymbol{\beta}_1 + \bar{\rho}\boldsymbol{\eta}_{2,k_2})\\
\mathbf{D}_{7,4} &:= \frac{\mathrm{d}\boldsymbol{\eta}_{2,k_2-1}}{\mathrm{d}\psi} = -\frac{1}{\sqrt{1+2\psi\rho+\psi^2}}\,(\mathbf{X}_1\boldsymbol{\beta}_1 + \bar{\rho}\boldsymbol{\eta}_{2,k_2-1})\\
\mathbf{D}_{7,5} &:= \frac{\mathrm{d}\boldsymbol{\eta}_3}{\mathrm{d}\psi} = \frac{1-\bar{\rho}^2}{\sqrt{1+2\psi\rho+\psi^2}}\,\mathbf{1}_n
\end{aligned}$$

where $\bar{\rho} := (\psi+\rho)/(\sqrt{1+2\psi\rho+\psi^2})$

- $\vartheta_8 = \tilde{\rho}$:

$$\begin{aligned}
\mathbf{D}_{8,3} &:= \frac{\mathrm{d}\boldsymbol{\eta}_{2,k_2}}{\mathrm{d}\tilde{\rho}} = -R\,\frac{\psi}{1+2\psi\rho+\psi^2}\,\boldsymbol{\eta}_{2,k_2} \\
\mathbf{D}_{8,4} &:= \frac{\mathrm{d}\boldsymbol{\eta}_{2,k_2-1}}{\mathrm{d}\tilde{\rho}} = -R\,\frac{\psi}{1+2\psi\rho+\psi^2}\,\boldsymbol{\eta}_{2,k_2-1} \\
\mathbf{D}_{8,5} &:= \frac{\mathrm{d}\boldsymbol{\eta}_3}{\mathrm{d}\tilde{\rho}} = R\left[\frac{1}{\sqrt{1+2\psi\rho+\psi^2}} - \frac{\psi\bar{\rho}}{1+2\psi\rho+\psi^2}\right]\mathbf{1}_n \\
R &:= \frac{\partial\rho}{\partial\tilde{\rho}} = \frac{4\exp\{2\tilde{\rho}\}}{(1+\exp\{2\tilde{\rho}\})^2}.
\end{aligned}$$

We further derive the elements of the vector $\mathbf{u}$ as follows:

- $\boldsymbol{\eta}_{1,k_1-l}$, $l = 0, 1$:

$$\begin{aligned}
\mathbf{u}_{1,1} &\equiv \frac{\mathrm{d}\Phi_2(\boldsymbol{\eta}_{1,k_1}, \boldsymbol{\eta}_{2,k_2}; \bar{\rho})}{\mathrm{d}\boldsymbol{\eta}_{1,k_1}} - \frac{\mathrm{d}\Phi_2(\boldsymbol{\eta}_{1,k_1}, \boldsymbol{\eta}_{2,k_2-1}; \bar{\rho})}{\mathrm{d}\boldsymbol{\eta}_{1,k_1}} \\
&= \phi(\boldsymbol{\eta}_{1,k_1})\left[\Phi\left(\frac{\boldsymbol{\eta}_{2,k_2} - \bar{\rho}\boldsymbol{\eta}_{1,k_1}}{\sqrt{1-\bar{\rho}^2}}\right) - \Phi\left(\frac{\boldsymbol{\eta}_{2,k_2-1} - \bar{\rho}\boldsymbol{\eta}_{1,k_1}}{\sqrt{1-\bar{\rho}^2}}\right)\right] \\
\mathbf{u}_{1,2} &\equiv \frac{\mathrm{d}\Phi_2(\boldsymbol{\eta}_{1,k_1-1}, \boldsymbol{\eta}_{2,k_2-1}; \bar{\rho})}{\mathrm{d}\boldsymbol{\eta}_{1,k_1-1}} - \frac{\mathrm{d}\Phi_2(\boldsymbol{\eta}_{1,k_1-1}, \boldsymbol{\eta}_{2,k_2}; \bar{\rho})}{\mathrm{d}\boldsymbol{\eta}_{1,k_1-1}} \\
&= \phi(\boldsymbol{\eta}_{1,k_1-1})\left[\Phi\left(\frac{\boldsymbol{\eta}_{2,k_2-1} - \bar{\rho}\boldsymbol{\eta}_{1,k_1-1}}{\sqrt{1-\bar{\rho}^2}}\right) - \Phi\left(\frac{\boldsymbol{\eta}_{2,k_2} - \bar{\rho}\boldsymbol{\eta}_{1,k_1-1}}{\sqrt{1-\bar{\rho}^2}}\right)\right]
\end{aligned}$$

- $\boldsymbol{\eta}_{2,k_2-m}$, $m = 0, 1$:

$$\begin{aligned}
\mathbf{u}_{2,1} &\equiv \frac{\mathrm{d}\Phi_2(\boldsymbol{\eta}_{1,k_1}, \boldsymbol{\eta}_{2,k_2}; \bar{\rho})}{\mathrm{d}\boldsymbol{\eta}_{2,k_2}} - \frac{\mathrm{d}\Phi_2(\boldsymbol{\eta}_{1,k_1-1}, \boldsymbol{\eta}_{2,k_2}; \bar{\rho})}{\mathrm{d}\boldsymbol{\eta}_{2,k_2}} \\
&= \phi(\boldsymbol{\eta}_{2,k_2})\left[\Phi\left(\frac{\boldsymbol{\eta}_{1,k_1} - \bar{\rho}\boldsymbol{\eta}_{2,k_2}}{\sqrt{1-\bar{\rho}^2}}\right) - \Phi\left(\frac{\boldsymbol{\eta}_{1,k_1-1} - \bar{\rho}\boldsymbol{\eta}_{2,k_2}}{\sqrt{1-\bar{\rho}^2}}\right)\right] \\
\mathbf{u}_{2,2} &\equiv \frac{\mathrm{d}\Phi_2(\boldsymbol{\eta}_{1,k_1-1}, \boldsymbol{\eta}_{2,k_2-1}; \bar{\rho})}{\mathrm{d}\boldsymbol{\eta}_{2,k_2-1}} - \frac{\mathrm{d}\Phi_2(\boldsymbol{\eta}_{1,k_1}, \boldsymbol{\eta}_{2,k_2-1}; \bar{\rho})}{\mathrm{d}\boldsymbol{\eta}_{2,k_2-1}} \\
&= \phi(\boldsymbol{\eta}_{2,k_2-1})\left[\Phi\left(\frac{\boldsymbol{\eta}_{1,k_1-1} - \bar{\rho}\boldsymbol{\eta}_{2,k_2-1}}{\sqrt{1-\bar{\rho}^2}}\right) - \Phi\left(\frac{\boldsymbol{\eta}_{1,k_1} - \bar{\rho}\boldsymbol{\eta}_{2,k_2-1}}{\sqrt{1-\bar{\rho}^2}}\right)\right]
\end{aligned}$$

- $\boldsymbol{\eta}_3$:

$$\begin{aligned}
\mathbf{u}_3 &= \sum_{l,m=\{0,1\}} (-1)^{l+m}\frac{\mathrm{d}\Phi_2(\boldsymbol{\eta}_{1,k_1-l}, \boldsymbol{\eta}_{2,k_2-m}; \bar{\rho})}{\mathrm{d}\boldsymbol{\eta}_3} \\
&= \frac{1}{2\pi\sqrt{1-\bar{\rho}^2}} \sum_{l,m=\{0,1\}} (-1)^{l+m} \exp\left\{\frac{\boldsymbol{\eta}_{1,k_1-l}^2 - 2\boldsymbol{\eta}_{1,k_1-l}\boldsymbol{\eta}_{2,k_2-m} + \boldsymbol{\eta}_{2,k_2-m}^2}{2(\bar{\rho}^2-1)}\right\}.
\end{aligned}$$

The Hessian matrix is made up of the components

$$\mathbf{W} = \widetilde{\mathbf{P}} \begin{bmatrix} \mathbf{W}_{1,1,1} & \mathbf{0}_{n,n} & \mathbf{W}_{1,1,3} & \mathbf{W}_{1,1,4} & \mathbf{W}_{1,1,5} \\ \bullet & \mathbf{W}_{1,2,2} & \mathbf{W}_{1,2,3} & \mathbf{W}_{1,2,4} & \mathbf{W}_{1,2,5} \\ \bullet & \bullet & \mathbf{W}_{1,3,3} & \mathbf{0}_{n,n} & \mathbf{W}_{1,3,5} \\ \bullet & \bullet & \bullet & \mathbf{W}_{1,4,4} & \mathbf{W}_{1,4,5} \\ \bullet & \bullet & \bullet & \bullet & \mathbf{W}_{1,5,5} \end{bmatrix} - \widetilde{\mathbf{P}}^2 \left[\mathbf{u}\mathbf{u}^\top \right]$$

and

$$\mathbf{K} = \begin{bmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \bullet & \mathbf{K}_{2,2} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \bullet & \bullet & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{K}_{3,7} & \mathbf{K}_{3,8} \\ \bullet & \bullet & \bullet & \mathbf{K}_{4,4} & \mathbf{0} & \mathbf{0} & \mathbf{K}_{4,7} & \mathbf{K}_{4,8} \\ \bullet & \bullet & \bullet & \bullet & \mathbf{0} & \mathbf{0} & \mathbf{K}_{5,7} & \mathbf{K}_{5,8} \\ \bullet & \bullet & \bullet & \bullet & \bullet & \mathbf{0} & \mathbf{K}_{6,7} & \mathbf{K}_{6,8} \\ \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \mathbf{K}_{7,7} & \mathbf{K}_{7,8} \\ \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \mathbf{K}_{8,8} \end{bmatrix},$$

with

- $\mathbf{u}_{1,1}$:

$$\begin{aligned} \mathbf{W}_{1,1,1} &\equiv \frac{\mathrm{d}\mathbf{u}_{1,1}}{\mathrm{d}\boldsymbol{\eta}_{1,k_1}} = \bar{\rho}(\mathbf{u}_{3,b} - \mathbf{u}_{3,a}) - \boldsymbol{\eta}_{1,k_1}\mathbf{u}_{1,1} \\ \mathbf{W}_{1,1,3} &\equiv \frac{\mathrm{d}\mathbf{u}_{1,1}}{\mathrm{d}\boldsymbol{\eta}_{2,k_2}} = \mathbf{u}_{3,a} \\ \mathbf{W}_{1,1,4} &\equiv \frac{\mathrm{d}\mathbf{u}_{1,1}}{\mathrm{d}\boldsymbol{\eta}_{2,k_2-1}} = -\mathbf{u}_{3,b} \\ \mathbf{W}_{1,1,5} &\equiv \frac{\mathrm{d}\mathbf{u}_{1,1}}{\mathrm{d}\boldsymbol{\eta}_3} = \frac{\bar{\rho}(\mathbf{u}_{3,a}\boldsymbol{\eta}_{2,k_2} - \mathbf{u}_{3,b}\boldsymbol{\eta}_{2,k_2-1}) + \boldsymbol{\eta}_{1,k_1}(\mathbf{u}_{3,b} - \mathbf{u}_{3,a})}{1 - \bar{\rho}^2} \end{aligned}$$

- $\mathbf{u}_{1,2}$:

$$\begin{aligned} \mathbf{W}_{1,2,2} &\equiv \frac{\mathrm{d}\mathbf{u}_{1,2}}{\mathrm{d}\boldsymbol{\eta}_{1,k_1-1}} = \bar{\rho}(\mathbf{u}_{3,c} - \mathbf{u}_{3,d}) - \boldsymbol{\eta}_{1,k_1-1}\mathbf{u}_{1,2} \\ \mathbf{W}_{1,2,3} &\equiv \frac{\mathrm{d}\mathbf{u}_{1,2}}{\mathrm{d}\boldsymbol{\eta}_{2,k_2}} = \mathbf{u}_{3,c} \\ \mathbf{W}_{1,2,4} &\equiv \frac{\mathrm{d}\mathbf{u}_{1,2}}{\mathrm{d}\boldsymbol{\eta}_{2,k_2-1}} = \mathbf{u}_{3,d} \\ \mathbf{W}_{1,2,5} &\equiv \frac{\mathrm{d}\mathbf{u}_{1,2}}{\mathrm{d}\boldsymbol{\eta}_3} = \frac{\bar{\rho}(\mathbf{u}_{3,d}\boldsymbol{\eta}_{2,k_2-1} - \mathbf{u}_{3,c}\boldsymbol{\eta}_{2,k_2}) + \boldsymbol{\eta}_{1,k_1-1}(\mathbf{u}_{3,c} - \mathbf{u}_{3,d})}{1 - \bar{\rho}^2} \end{aligned}$$

- $\mathbf{u}_{2,1}$:

$$\begin{aligned}\mathbf{W}_{1,3,3} &\equiv \frac{\mathrm{d}\mathbf{u}_{2,1}}{\mathrm{d}\boldsymbol{\eta}_{1,k_2}} = \bar{\rho}(\mathbf{u}_{3,c} - \mathbf{u}_{3,a}) - \boldsymbol{\eta}_{2,k_2}\mathbf{u}_{2,1} \\ \mathbf{W}_{1,3,5} &\equiv \frac{\mathrm{d}\mathbf{u}_{2,1}}{\mathrm{d}\boldsymbol{\eta}_3} = \frac{\bar{\rho}(\mathbf{u}_{3,a}\boldsymbol{\eta}_{1,k_1} - \mathbf{u}_{3,c}\boldsymbol{\eta}_{1,k_1-1}) + \boldsymbol{\eta}_{2,k_2}(\mathbf{u}_{3,c} - \mathbf{u}_{3,a})}{1-\bar{\rho}^2}\end{aligned}$$

- $\mathbf{u}_{2,2}$:

$$\begin{aligned}\mathbf{W}_{1,4,4} &\equiv \frac{\mathrm{d}\mathbf{u}_{2,2}}{\mathrm{d}\boldsymbol{\eta}_{1,k_2-1}} = \bar{\rho}(\mathbf{u}_{3,b} - \mathbf{u}_{3,d}) - \boldsymbol{\eta}_{2,k_2-1}\mathbf{u}_{2,2} \\ \mathbf{W}_{1,4,5} &\equiv \frac{\mathrm{d}\mathbf{u}_{2,2}}{\mathrm{d}\boldsymbol{\eta}_3} = \frac{\bar{\rho}(\mathbf{u}_{3,d}\boldsymbol{\eta}_{1,k_1-1} - \mathbf{u}_{3,b}\boldsymbol{\eta}_{1,k_1}) + \boldsymbol{\eta}_{2,k_2-1}(\mathbf{u}_{3,b} - \mathbf{u}_{3,d})}{1-\bar{\rho}^2}\end{aligned}$$

- $\mathbf{u}_3$:

$$\begin{aligned}\mathbf{W}_{1,5,5} \equiv \frac{\mathrm{d}\mathbf{u}_3}{\mathrm{d}\boldsymbol{\eta}_3} = \frac{1}{(1-\bar{\rho}^2)^2} \sum_{l,m=\{0,1\}} & \mathbf{u}_{3,(l,m)} \\ & \left[\boldsymbol{\eta}_{1,k_1-l}\boldsymbol{\eta}_{2,k_2-m}(1+\bar{\rho}^2) - \bar{\rho}\,(\boldsymbol{\eta}_{1,k_1-l}^2 + \boldsymbol{\eta}_{2,k_2-m}^2 + \bar{\rho}^2 - 1)\right]\end{aligned}$$

where $\mathbf{u}_{3,(l,m)}$ represents the (l,m)th addendum defining $\mathbf{u}_3$.

Concerning $\mathbf{K}$, we derive:

- $\boldsymbol{\vartheta}^\top = \tilde{\boldsymbol{c}}_{1,h}^\top$:

$$\begin{aligned}\mathbf{K}_{2,2,a} &= \frac{\mathrm{d}^2\boldsymbol{\eta}_{1,k_1}}{\mathrm{d}\tilde{\boldsymbol{c}}_{1,h'}\mathrm{d}\tilde{\boldsymbol{c}}_{1,h}^\top} = 2\;\mathbb{1}_{k_1\geq h}\mathbb{1}_{h=h'} \\ \mathbf{K}_{2,2,b} &= \frac{\mathrm{d}^2\boldsymbol{\eta}_{1,k_1-1}}{\mathrm{d}\tilde{\boldsymbol{c}}_{1,h'}\mathrm{d}\tilde{\boldsymbol{c}}_{1,h}^\top} = 2\;\mathbb{1}_{k_1\geq h+1}\mathbb{1}_{h=h'} \\ \mathbf{K}_{2,2} &= \mathbf{P}(\mathbf{u}_{1,1}\mathbf{K}_{2,2,a} + \mathbf{u}_{1,2}\mathbf{K}_{2,2,b})\end{aligned}$$

- $\vartheta = \tilde{c}_{2,1}$:

$$\begin{aligned}\mathbf{K}_{3,7,a} &= \frac{\mathrm{d}^2\boldsymbol{\eta}_{2,k_2}}{\mathrm{d}\psi\mathrm{d}\tilde{c}_{2,1}} = -\frac{\bar{\rho}}{\sqrt{1+2\psi\rho+\psi^2}}\,\mathbf{D}_{3,3} \\ \mathbf{K}_{3,7,b} &= \frac{\mathrm{d}^2\boldsymbol{\eta}_{2,k_2-1}}{\mathrm{d}\psi\mathrm{d}\tilde{c}_{2,1}} = -\frac{\bar{\rho}}{\sqrt{1+2\psi\rho+\psi^2}}\,\mathbf{D}_{3,4} \\ \mathbf{K}_{3,7} &= \mathbf{P}(\mathbf{u}_{2,1}\mathbf{K}_{3,7,a} + \mathbf{u}_{2,2}\mathbf{K}_{3,7,b})\end{aligned}$$

$$\begin{aligned}
\mathbf{K}_{3,8,a} &= \frac{\mathrm{d}^2\boldsymbol{\eta}_{2,k_2}}{\mathrm{d}\tilde{\rho}\mathrm{d}\tilde{c}_{2,1}} = -\left[\frac{\psi}{1+2\psi\rho+\psi^2}\,\mathbf{D}_{3,3}\right]R \\
\mathbf{K}_{3,8,b} &= \frac{\mathrm{d}^2\boldsymbol{\eta}_{2,k_2-1}}{\mathrm{d}\tilde{\rho}\mathrm{d}\tilde{c}_{2,1}} = -\left[\frac{\psi}{1+2\psi\rho+\psi^2}\,\mathbf{D}_{3,4}\right]R \\
\mathbf{K}_{3,8} &= \mathbf{P}(\mathbf{u}_{2,1}\mathbf{K}_{3,8,a} + \mathbf{u}_{2,2}\mathbf{K}_{3,8,b})
\end{aligned}$$

- $\boldsymbol{\vartheta}^\top = \tilde{\boldsymbol{c}}_{2,h}^\top$:

$$\begin{aligned}
\mathbf{K}_{4,4,a} &= \frac{\mathrm{d}^2\boldsymbol{\eta}_{2,k_2}}{\mathrm{d}\tilde{\boldsymbol{c}}_{2,h'}\mathrm{d}\tilde{\boldsymbol{c}}_{2,h}^\top} = 2\,\frac{\mathbb{1}_{k_2\geq h}\mathbb{1}_{h=h'}}{\sqrt{1+2\psi\rho+\psi^2}} \\
\mathbf{K}_{4,4,b} &= \frac{\mathrm{d}^2\boldsymbol{\eta}_{2,k_2-1}}{\mathrm{d}\tilde{\boldsymbol{c}}_{2,h'}\mathrm{d}\tilde{\boldsymbol{c}}_{2,h}^\top} = 2\,\frac{\mathbb{1}_{k_2\geq h+1}\mathbb{1}_{h=h'}}{\sqrt{1+2\psi\rho+\psi^2}} \\
\mathbf{K}_{4,4} &= \mathbf{P}(\mathbf{u}_{2,1}\mathbf{K}_{4,4,a} + \mathbf{u}_{2,2}\mathbf{K}_{4,4,b})
\end{aligned}$$

$$\begin{aligned}
\mathbf{K}_{4,7,a} &= \frac{\mathrm{d}^2\boldsymbol{\eta}_{2,k_2}}{\mathrm{d}\psi\mathrm{d}\tilde{\boldsymbol{c}}_{2,h}^\top} = -\frac{\bar{\rho}}{\sqrt{1+2\psi\rho+\psi^2}}\,\mathbf{D}_{4,3} \\
\mathbf{K}_{4,7,b} &= \frac{\mathrm{d}^2\boldsymbol{\eta}_{2,k_2-1}}{\mathrm{d}\psi\mathrm{d}\tilde{\boldsymbol{c}}_{2,h}^\top} = -\frac{\bar{\rho}}{\sqrt{1+2\psi\rho+\psi^2}}\,\mathbf{D}_{4,4} \\
\mathbf{K}_{4,7} &= \mathbf{P}(\mathbf{u}_{2,1}\mathbf{K}_{4,7,a} + \mathbf{u}_{2,2}\mathbf{K}_{4,7,b})
\end{aligned}$$

$$\begin{aligned}
\mathbf{K}_{4,8,a} &= \frac{\mathrm{d}^2\boldsymbol{\eta}_{2,k_2}}{\mathrm{d}\tilde{\rho}\mathrm{d}\tilde{\boldsymbol{c}}_{2,h}^\top} = -\left[\frac{\psi}{1+2\psi\rho+\psi^2}\,\mathbf{D}_{4,3}\right]R \\
\mathbf{K}_{4,8,b} &= \frac{\mathrm{d}^2\boldsymbol{\eta}_{2,k_2-1}}{\mathrm{d}\tilde{\rho}\mathrm{d}\tilde{\boldsymbol{c}}_{2,h}^\top} = -\left[\frac{\psi}{1+2\psi\rho+\psi^2}\,\mathbf{D}_{4,4}\right]R \\
\mathbf{K}_{4,8} &= \mathbf{P}(\mathbf{u}_{2,1}\mathbf{K}_{4,8,a} + \mathbf{u}_{2,2}\mathbf{K}_{4,8,b})
\end{aligned}$$

- $\boldsymbol{\vartheta}^\top = \boldsymbol{\beta}_1^\top$:

$$\begin{aligned}
\mathbf{K}_{5,7,a} &= \frac{\mathrm{d}^2\boldsymbol{\eta}_{2,k_2}}{\mathrm{d}\psi\mathrm{d}\boldsymbol{\beta}_1^\top} = \frac{\mathrm{d}^2\boldsymbol{\eta}_{2,k_2-1}}{\mathrm{d}\psi\mathrm{d}\boldsymbol{\beta}_1^\top} = -\frac{\mathbf{X}_1 + \mathbf{D}_{5,3}\bar{\rho}}{\sqrt{1+2\psi\rho+\psi^2}} \\
\mathbf{K}_{5,7} &= \mathbf{P}(\mathbf{u}_{2,1}\mathbf{K}_{5,7,a} + \mathbf{u}_{2,2}\mathbf{K}_{5,7,b})
\end{aligned}$$

$$\begin{aligned}\mathbf{K}_{5,8,a} &= \frac{\mathrm{d}^2\boldsymbol{\eta}_{2,k_2}}{\mathrm{d}\tilde{\rho}\mathrm{d}\boldsymbol{\beta}_1^\top} = \frac{\mathrm{d}^2\boldsymbol{\eta}_{2,k_2-1}}{\mathrm{d}\tilde{\rho}\mathrm{d}\boldsymbol{\beta}_1^\top} = -\left[\frac{\psi}{1+2\psi\rho+\psi^2}\,\mathbf{D}_{5,3}\right]R \\ \mathbf{K}_{5,8} &= \mathbf{P}(\mathbf{u}_{2,1}\mathbf{K}_{5,8,a} + \mathbf{u}_{2,2}\mathbf{K}_{5,8,b})\end{aligned}$$

- $\boldsymbol{\vartheta}^\top = \boldsymbol{\beta}_2^\top$:

$$\begin{aligned}\mathbf{K}_{6,7,a} &= \frac{\mathrm{d}^2\boldsymbol{\eta}_{2,k_2}}{\mathrm{d}\psi\mathrm{d}\boldsymbol{\beta}_2^\top} = \frac{\mathrm{d}^2\boldsymbol{\eta}_{2,k_2-1}}{\mathrm{d}\psi\mathrm{d}\boldsymbol{\beta}_2^\top} = -\frac{\bar{\rho}}{\sqrt{1+2\psi\rho+\psi^2}}\,\mathbf{D}_{6,3} \\ \mathbf{K}_{6,7} &= \mathbf{P}(\mathbf{u}_{2,1}\mathbf{K}_{5,7,a} + \mathbf{u}_{2,2}\mathbf{K}_{5,7,b})\end{aligned}$$

$$\begin{aligned}\mathbf{K}_{6,8,a} &= \frac{\mathrm{d}^2\boldsymbol{\eta}_{2,k_2}}{\mathrm{d}\tilde{\rho}\mathrm{d}\boldsymbol{\beta}_2^\top} = \frac{\mathrm{d}^2\boldsymbol{\eta}_{2,k_2-1}}{\mathrm{d}\tilde{\rho}\mathrm{d}\boldsymbol{\beta}_2^\top} = -\left[\frac{\psi}{1+2\psi\rho+\psi^2}\,\mathbf{D}_{6,3}\right]R \\ \mathbf{K}_{6,8} &= \mathbf{P}(\mathbf{u}_{2,1}\mathbf{K}_{5,8,a} + \mathbf{u}_{2,2}\mathbf{K}_{5,8,b})\end{aligned}$$

- $\vartheta = \psi$:

$$\begin{aligned}\mathbf{K}_{7,7,a} &= \frac{\mathrm{d}^2\boldsymbol{\eta}_{2,k_2}}{\mathrm{d}\psi^2} = -\frac{\mathbf{D}_{7,5}\boldsymbol{\eta}_{2,k_2} + 2\mathbf{D}_{7,3}\bar{\rho}}{\sqrt{1+2\psi\rho+\psi^2}} \\ \mathbf{K}_{7,7,b} &= \frac{\mathrm{d}^2\boldsymbol{\eta}_{2,k_2-1}}{\mathrm{d}\psi^2} = -\frac{\mathbf{D}_{7,5}\boldsymbol{\eta}_{2,k_2-1} + 2\mathbf{D}_{7,4}\bar{\rho}}{\sqrt{1+2\psi\rho+\psi^2}} \\ \mathbf{K}_{7,7,c} &= \frac{\mathrm{d}^2\boldsymbol{\eta}_3}{\mathrm{d}\psi^2} = -\frac{2\mathbf{D}_{7,5}\bar{\rho}\sqrt{1+2\psi\rho+\psi^2} + \bar{\rho}(1-\bar{\rho}^2)}{1+2\psi\rho+\psi^2} \\ \mathbf{K}_{7,7} &= \mathbf{P}(\mathbf{u}_{2,1}\mathbf{K}_{7,7,a} + \mathbf{u}_{2,2}\mathbf{K}_{7,7,b} + \mathbf{u}_3\mathbf{K}_{7,7,c})\end{aligned}$$

$$\begin{aligned}\mathbf{K}_{7,8,a} &= \frac{\mathrm{d}^2\boldsymbol{\eta}_{2,k_2}}{\mathrm{d}\tilde{\rho}\mathrm{d}\psi} = -\frac{1}{1+2\psi\rho+\psi^2}\left[\left(1-2\frac{\psi\bar{\rho}}{\sqrt{1+2\psi\rho+\psi^2}}\right)\boldsymbol{\eta}_{2,k_2} + \mathbf{D}_{7,3}\psi\right]R \\ \mathbf{K}_{7,8,b} &= \frac{\mathrm{d}^2\boldsymbol{\eta}_{2,k_2}}{\mathrm{d}\tilde{\rho}\mathrm{d}\psi} \\ &= -\frac{1}{1+2\psi\rho+\psi^2}\left[\left(1-2\frac{\psi\bar{\rho}}{\sqrt{1+2\psi\rho+\psi^2}}\right)\boldsymbol{\eta}_{2,k_2-1} + \mathbf{D}_{7,4}\psi\right]R \\ \mathbf{K}_{7,8,c} &= \frac{\mathrm{d}^2\boldsymbol{\eta}_{2,k_2}}{\mathrm{d}\tilde{\rho}\mathrm{d}\psi} = -\frac{1}{1+2\psi\rho+\psi^2}\left[\left(1-2\frac{\psi\bar{\rho}}{\sqrt{1+2\psi\rho+\psi^2}}\right)2\bar{\rho} + \mathbf{D}_{7,5}\psi\right]R \\ \mathbf{K}_{7,8} &= \mathbf{P}(\mathbf{u}_{2,1}\mathbf{K}_{7,8,a} + \mathbf{u}_{2,2}\mathbf{K}_{7,8,b} + \mathbf{u}_3\mathbf{K}_{7,8,c})\end{aligned}$$

- $\vartheta = \tilde{\rho}$:

$$
\begin{aligned}
\mathbf{K}_{8,8,a} &= \frac{\mathrm{d}^2\boldsymbol{\eta}_{2,k_2}}{\mathrm{d}\tilde{\rho}^2} = \mathbf{D}_{8,3}\left[\frac{R2}{R} - 3\frac{\psi}{1+2\psi\rho+\psi^2}\, R\right] \\
\mathbf{K}_{8,8,b} &= \frac{\mathrm{d}^2\boldsymbol{\eta}_{2,k_2-1}}{\mathrm{d}\tilde{\rho}^2} = \mathbf{D}_{8,4}\left[\frac{R2}{R} - 3\frac{\psi}{1+2\psi\rho+\psi^2}\, R\right] \\
\mathbf{K}_{8,8,c} &= \frac{\mathrm{d}^2\boldsymbol{\eta}_3}{\mathrm{d}\tilde{\rho}^2} = \left(R2 - 2\frac{\psi}{1+2\psi\rho+\psi^2}\, R^2\right) \\
&\qquad \left(\frac{1}{\sqrt{1+2\psi\rho+\psi^2}} - \frac{\psi\bar{\rho}}{1+2\psi\rho+\psi^2}\right) + \bar{\rho}\left(\frac{\psi}{1+2\psi\rho+\psi^2}\right)^2 \\
R2 &\equiv \frac{\mathrm{d}R}{\mathrm{d}\tilde{\rho}} = \frac{\mathrm{d}}{\mathrm{d}\tilde{\rho}}\left[\frac{4\exp\{2\tilde{\rho}\}}{(1+\exp\{2\tilde{\rho}\})^2}\right] = \frac{8\exp\{2\tilde{\rho}\}(1-\exp\{2\tilde{\rho}\})}{(1+\exp\{2\tilde{\rho}\})^3} \\
\mathbf{K}_{8,8} &= \mathbf{P}(\mathbf{u}_{2,1}\mathbf{K}_{8,8,a} + \mathbf{u}_{2,2}\mathbf{K}_{8,8,b} + \mathbf{u}_3\mathbf{K}_{8,8,c})
\end{aligned}
$$

B.1.2 Some Simulation Evidence

The simulated scenario comprises a bivariate system of equations specified by the following Data Generating Process (DGP):

$$
\begin{aligned}
\mathrm{y}^*_{1,i} &= \mathrm{x}_{1,i} + 2\mathrm{x}_{2,i} + \mathrm{x}_{3,i} + s_{1,1}(\mathrm{v}_{1,i}) + s_{1,2}(\mathrm{v}_{2,i}) + \varepsilon_1 \\
\mathrm{y}^*_{2,i} &= -0.3\mathrm{y}^*_{1,i} + \mathrm{x}_{1,i} - 2\mathrm{x}_{2,i} + s_{2,1}(\mathrm{v}_{1,i}) + \varepsilon_2
\end{aligned}
\qquad
\boldsymbol{\varepsilon}_i \sim \mathcal{N}_2\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right)
$$

for various specifications of the correlation coefficient, $\rho = \{0.1, 0.5, 0.9\}$, indicating an increasing strength of the unmeasured confounding problem. The test functions are displayed in red in Figure 2.1 in the main text, and given by $s_{1,1}(\mathrm{v}_{1,i}) = -0.7\{4\mathrm{v}_{1,i} + 2.5\mathrm{v}^2_{1,i} + 0.7\sin(5\mathrm{v}_{1,i}) + \cos(7.5\mathrm{v}_{1,i})\}$, $s_{1,2}(\mathrm{v}_{2,i}) = -0.4\{-0.3 - 1.6\mathrm{v}_{2,i} + \sin(5\mathrm{v}_{2,i})\}$ and $s_{2,1}(\mathrm{v}_{1,i}) = 0.6\{\exp\{\mathrm{v}_{1,i}\} + \sin(2,9\mathrm{v}_{1,i})\}$. Furthermore, the ordered values of $y_{j,i}$ have been computed following the observation rule:

$$
\mathrm{y}_{j,i} = \sum_{k_j \in \mathcal{K}_j} k_j \mathbb{1}_{c_{j,k_j-1} < \mathrm{y}^*_{j,i} \leq c_{j,k_j}},
$$

for every $j \in \{1, 2\}$, and obtained by setting the threshold parameters at $\boldsymbol{c}_1 := (-2, -1, 0, 2)^\top$ and $\boldsymbol{c}_2 := (-7, -2, -1, 1, 2)^\top$. Finally, sample sizes are specified at 500, 3,000, and 10,000, and $N = 100$ replications of each design are performed: in conclusion, this experiment encompassed 9 different simulated scenarios whose results are detailed below. In particular, we focus on the sampling properties of $\widehat{\psi}$ and $\widehat{\rho}$, namely the parameters that are neglected whenever a naive ordered regression model is fitted.

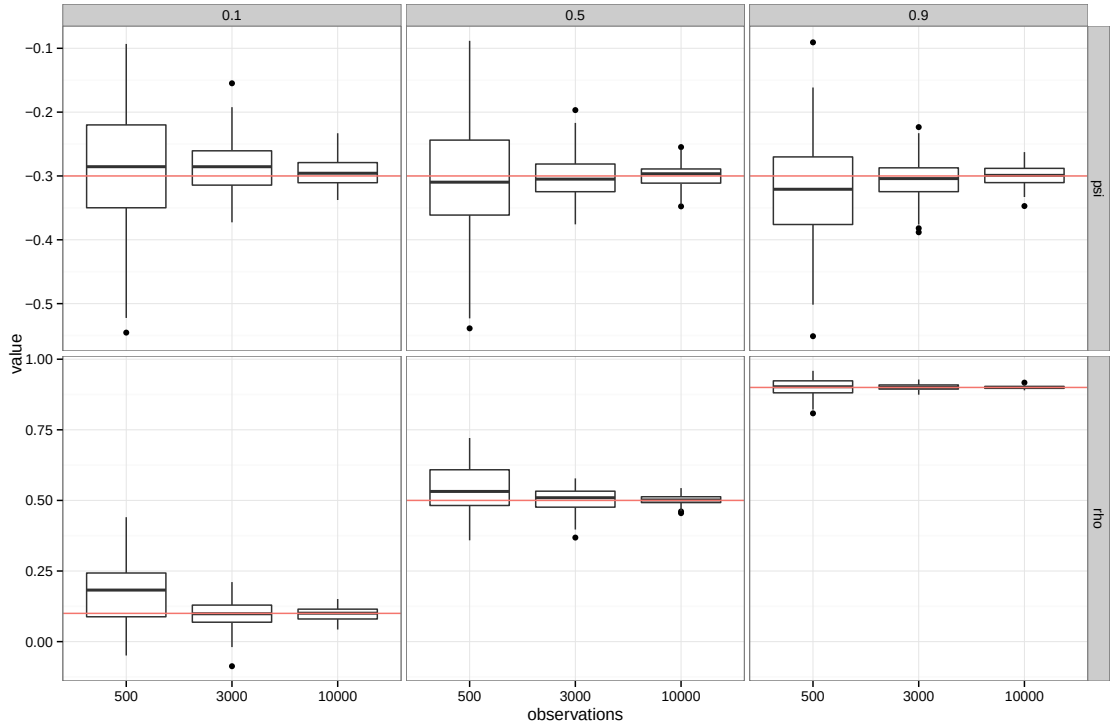

Figure B.1: Box plots corresponding to the estimates of ψ and ρ for different sample sizes and correlation coefficients (0.1, 0.5 and 0.9) where the true values are denoted by a red line in each of the panels. Results are obtained using 100 replications of the DGP detailed in this section.

In general, the simulations show (Figure B.1) that the parameters of interest are unbiased starting from moderate sample sizes (i.e. 3,000) while, for $n = 500$, our algorithm produced estimates for the correlation coefficient quite far from the corresponding simulated values: this tendency is nonetheless overcome once the number of observations is increased. We also stress that, as n increases, the estimated parameter vector approaches its true value with a lower standard deviation and, interestingly, $\widehat{\rho}$ is on average closer to its simulated values at any given sample size and higher magnitude. This fact is not unexpected, and similar occurrences have been also reported by Chib and Greenberg (2007) and Marra and Radice (2011) in the context of a semi-parametric regression for dichotomous responses. In particular, they explained the situation by noticing that a higher correlation coefficient is a signal of a more severe role of the unobservables in determining the association between Y_1 and Y_2 which is more easily measured by a bivariate model. Stated differently, the joint estimation of the parameters in the model allows for the full use of all the information contained in the data.

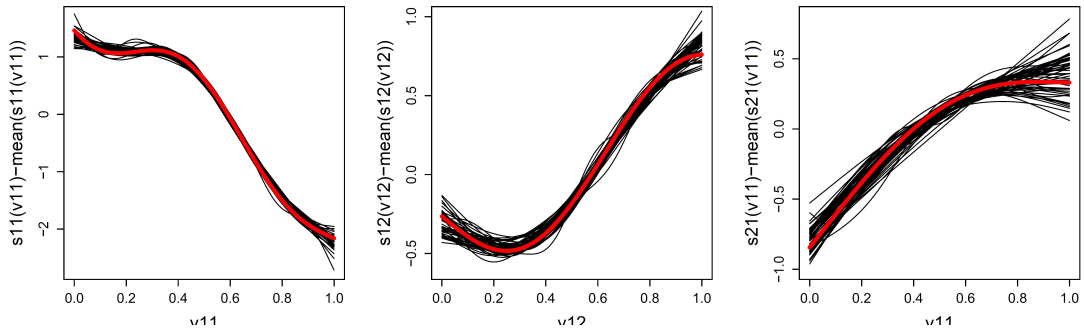

Figure B.2: Estimated smooth curves obtained from 50 replicates of a Monte Carlo experiment comprising 3,000 simulated observations (true curves in red). The DGP is given in this section, and ψ and ρ set to -0.3 and 0.2, respectively. Refer to the caption of Figure 2.1 for more details.

B.1.3 Further Details on the Empirical Illustration

Variable	BCS70	Definition	Levels
`edu`	HIACA00	Respondent's highest education achieved	$= 01$: $\leq$ O-levels
			$= 02$: A-levels
			$= 03$: Higher Education
`drk`	drinking	Drinking frequency	$= 01$: special occasions
			$= 02$: 2-3 times/month
			$= 03$: once a week
			$= 04$: 2-3 days/week
			$= 05$: most days
`drk5`	drinking, beer, spirits, wine, pops,sherry	Alcohol intake	$= 01$: special occasions
			$= 02$: 2-3 times/month
			$= 03$: $<$ NHS limits
			$= 04$: $\approx$ NHS limits
			$= 05$: $>$ NHS limits
`BAS.tot`	i3504-i3644	Total reported BAS (`edu`)	– continuous
`region`	BD3REGN	Region of residence (`drk5`/`drk`)	$= 01$: South East
			$= 02$: Scotland
			$= 03$: Wales
			$= 04$: South West
			$= 05$: missing/unknown
			$= 06$: East Anglia
			$= 07$: West Midlands
			$= 08$: East Midlands
			$= 09$: Yorks & Humber.
			$= 10$: North West
			$= 11$: North

Table B.1: Description of the responses and the equation-specific covariates included in the study. The dependent variable `drk5` has been obtained by replacing levels 03-05 of `drk` by an equivalent (averaged) amount of alcohol units as based on the following conversion: 1 pint of beer: 2.8u; 1 glass of spirits/sherry: 1u; 1 glass of wine: 2.1u; 1 bottle of alcopop: 1.4u.

Variable	BCS70	Definition	Levels
mum.not.pres.	a5.1	Relationship of mother figure	= 1: other = 2: missing = 3: natural mother
dad.not.pres.	a6.1	Relationship of father figure	= 1: other = 2: missing = 3: natural father
mum.edu	c1.12-c1.20	Mother's qualification	= 1: < O-levels = 2: missing = 3: O-levels = 4: A-levels = 5: professional = 6: degree
dad.edu	c1.1-c1.11	Father's qualification	= 1: < O-levels = 2: missing = 3: O-levels = 4: A-levels = 5: professional = 6: degree
s.class	BD3PSOC	Social class from father's occupation	= 1: professional = 2: missing/no data = 3: manual & tech. = 4: partially skilled = 5: non-manual = 6: manual = 7: unskilled
eth.child	a12.1	Ethnic group	= 1: white European = 2: missing = 3: other
mum.int.edu	j097	Mother's interest in child's education	= 1: very/moderate = 2: other = 3: cannot say = 4: no/very little
dad.int.edu	j098	Father's interest in child's education	= 1: very/moderate = 2: other = 3: cannot say = 4: no/very little
sex.b	sex10	Gender	= 0: girl = 1: boy
home	d2	Home tenure	= 1: owned outright = 2: missing = 3: rented/other = 4: being bought
mum.wrk.hr	c5.1	Mother's weekly working hours	– continuous

Table B.2: Description of the covariates that are common to both equations.

Highest Education	Drinking Frequency				
	no/occasional	light	1/week	2-3/week	most days
Up to O-levels	0.2028	0.1455	0.2256	0.3179	0.1081
	(.1294; .2941)	(.0535; .2203)	(.2154; .2358)	(.3090; .3263)	(.1020; .1139)
A-levels	0.1621	0.1328	0.2225	0.3480	0.1345
	(.0990; .2441)	(.0498; .1972)	(.2124; .2324)	(.3388; .3571)	(.1279; .1413)
HE or equivalent	0.1426	0.1237	0.2159	0.3615	0.1563
	(.0853; .2188)	(.0469; .1816)	(.2063; .2256)	(.3531; .3699)	(.1516; .1609)

Table B.3: Average predicted conditional probabilities when a measure of drinking frequency (`drk`) is used as response variable. More details are reported in the caption of Table 2.4.

Highest Education	Alcohol Consumption				
			at least one drink per week		
	no/occasional	light	$<$ NHS limits	$\approx$ NHS limits	$>$ NHS limits
Up to O-levels	0.2320	0.1424	0.2550	0.0926	0.2780
	(.1525; .3285)	(.0402; .2242)	(.2447; .2648)	(.0839; .1013)	(.2699; .2856)
A-levels	0.2029	0.1354	0.2554	0.0967	0.3095
	(.1299; .2944)	(.0388; .2105)	(.2450; .2651)	(.0877; .1058)	(.3012; .3177)
HE or equivalent	0.1876	0.1303	0.2530	0.0983	0.3308
	(.1184; .2772)	(.0378; .1998)	(.2427; .2626)	(.0889; .1075)	(.3242; .3373)

Table B.4: Average predicted conditional probabilities when the covariate `mum.wrk.hr` is dropped from the first equation. More details are reported in the caption of Table 2.4.

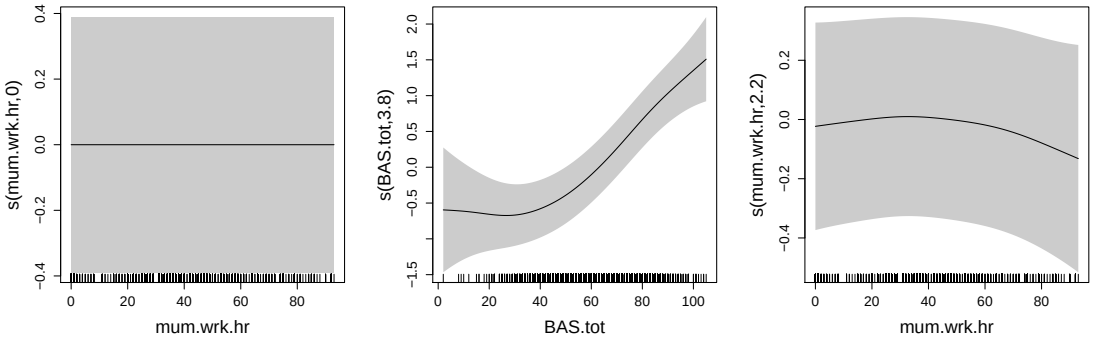

Figure B.3: Shrinkage method applied to the model specification of Section 2.4.1: `mum.wrk.hr` is not an influential predictor for children's education achievements.

B.2 Complements to Chapter 3

B.2.1 Analytical Definition of Bivariate Copulae

Any 2-dimensional copula is a function $\mathcal{C}_2$ with domain $[0,1]^2$ such that: (i) $\mathcal{C}_2$ is grounded and 2-increasing, and (ii) has margins $F_{1,j}$, $j \in \{1,2\}$, which satisfy $F_{1,j}(u_j) = u_j$ for all $u_j \in [0,1]$.

Let $(u_1, u_2) \in [0,1]^2$, and $\mathcal{C}_2(u_1, u_2) = 0$ whenever $u_j = 0$ for at least one j; then $\mathcal{C}_2$ is said to be *grounded*. Moreover, we define

$$B := [\boldsymbol{a}, \boldsymbol{b}] := [a_1, b_1] \times [a_2, b_2] \quad a_j \leq b_j \ \ \forall j$$

to be the 2-box whose vertices are in $[0,1]^2$, and

$$V_{\mathcal{C}_2}(B) := \Delta_{a_2}^{b_2} \Delta_{a_1}^{b_1} \mathcal{C}_2(u_1, u_2)$$

the corresponding $\mathcal{C}_2$-volume of B, with

$$\Delta_{a_1}^{b_1} \mathcal{C}_2(u_1, u_2) := \mathcal{C}_2(b_1, u_2) - \mathcal{C}_2(a_1, u_2) \quad \text{and} \quad \Delta_{a_2}^{b_2} \mathcal{C}_2(u_1, u_2) := \mathcal{C}_2(u_1, b_2) - \mathcal{C}_2(u_1, a_2)$$

being first order differences. We say that the real function $\mathcal{C}_2$ is 2-*increasing* if $V_{\mathcal{C}_2}(B) \geq 0$ for all 2-boxes whose vertices lie in $[0,1]^2$.

B.2.2 Copula Rotations

Rotated copulae can be obtained by applying the following transformations:

$$\begin{aligned} \mathcal{C}_{90}(u_1, u_2) &= u_1 - \mathcal{C}_\gamma(1 - u_1, u_2) \\ \mathcal{C}_{180}(u_1, u_2) &= u_1 + u_2 - 1 + \mathcal{C}_\gamma(1 - u_1, 1 - u_2) \\ \mathcal{C}_{270}(u_1, u_2) &= u_1 - \mathcal{C}_\gamma(u_1, 1 - u_2), \end{aligned}$$

where we have followed the convention of labelling the marginals corresponding to $j = 1, 2$ with u_1 and u_2, respectively. Contour plots of the copuae implemented in `CopulaCLM` are given in Figure B.4.

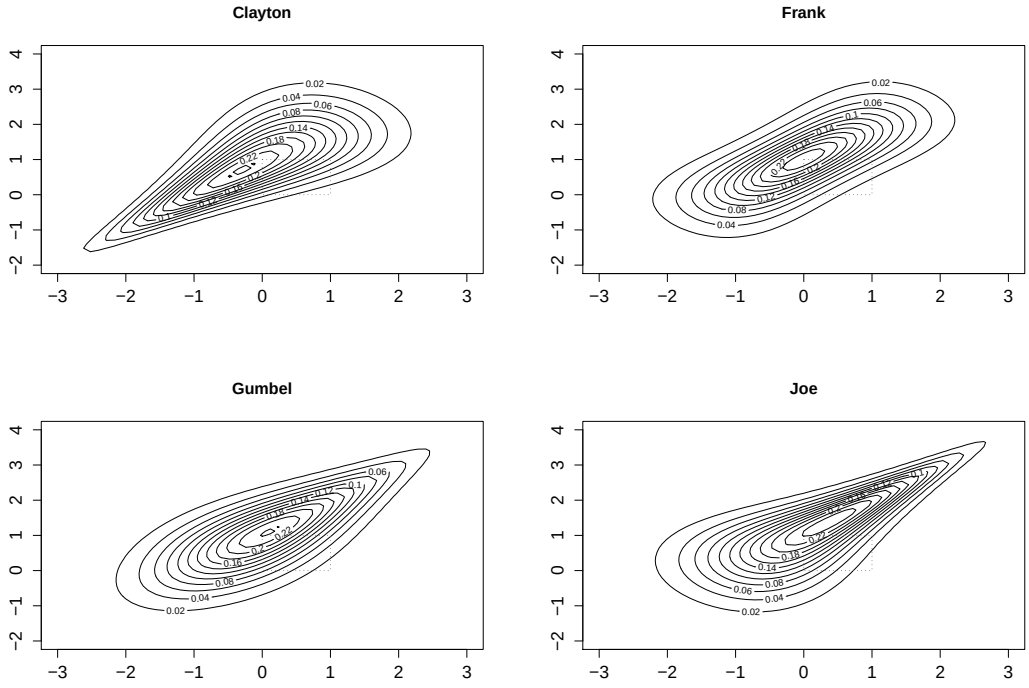

Figure B.4: Contour plots of some of the copula functions with standard normal margins for data simulated using association parameters γ of 2, 5.74, 2 and 2.86, respectively (these values are consistent with a medium positive correlation). The Frank copula allows for equal degrees of positive and negative dependence, whereas Clayton is asymmetric with a strong lower tail dependence but a weaker upper tail dependence. Vice versa for the Gumbel and Joe copulas.

B.2.3 Data Generating Processes Employed in Simulations

DGP for Figure 3.2

$$\begin{aligned} \mathrm{y}^*_{1,i} &= \mathrm{x}_{1,i} + 2\mathrm{x}_{2,i} + \mathrm{x}_{3,i} + s_{1,1}(\mathrm{v}_{1,i}) + s_{1,2}(\mathrm{v}_{2,i}) + \varepsilon_1 \\ \mathrm{y}^*_{2,i} &= 2\mathrm{x}_{1,i} - 2\mathrm{x}_{2,i} + s_{2,1}(\mathrm{v}_{1,i}) + \varepsilon_2 \end{aligned} \qquad \varepsilon_j \sim \mathcal{N}(0,1).$$

The test functions are given by

$$\begin{aligned} s_{1,1}(\mathrm{v}_{1,i}) &= -0.7\{4\mathrm{v}_{1,i} + 2.5\mathrm{v}_{1,i}^2 + 0.7\sin(5\mathrm{v}_{1,i}) + \cos(7.5\mathrm{v}_{1,i})\} \\ s_{1,2}(\mathrm{v}_{2,i}) &= -0.4\{-0.3 - 1.6\mathrm{v}_{2,i} + \sin(5\mathrm{v}_{2,i})\} \\ s_{2,1}(\mathrm{v}_{1,i}) &= 0.6\{\exp\{\mathrm{v}_{1,i}\} + \sin(2,9\mathrm{v}_{1,i})\}, \end{aligned}$$

and the ordered values of $\mathrm{y}_{j,i}$ have been computed following the observation rule

$$\mathrm{y}_{j,i} = \sum_{k_j \in \mathcal{K}_j} k_j \mathbb{1}_{c_{j,k_j-1} < \mathrm{y}^*_{j,i} \leq c_{j,k_j}}$$

for every $j \in \{1,2\}$, and obtained by setting the threshold parameters at $\boldsymbol{c}_{1,k} := (-2,-1,0,2)^\top$ and $\boldsymbol{c}_{2,k} := (-1.4,-0.7,-0.2,0.7,3)^\top$. Furthermore, we have set the copula association parameters at $\gamma_{\text{clayton}} = 0.2222$, $\gamma_{\text{frank}} = 0.9074$ and $\gamma_{\text{joe}} = 1.1944$, all corresponding to a Kendall's Tau of 0.1.

DGP for Figure 3.3 The same as of the previous paragraph but with smooth curves

$$\begin{aligned}
s_{1,1}(\mathrm{v}_{1,i}) &= 1 - \mathrm{v}_{1,i} + 1.6\mathrm{v}_{1,i}^4 - \sin(5\mathrm{v}_{1,i}) \\
s_{1,2}(\mathrm{v}_{2,i}) &= 4\mathrm{v}_{2,i} \\
s_{2,1}(\mathrm{v}_{1,i}) &= 0.08\{\mathrm{v}_{1,i}^{11}[10(1-\mathrm{v}_{1,i})]^6\} + 10(10\mathrm{v}_{1,i})(1-\mathrm{v}_{1,i})^{10},
\end{aligned}$$

and cut points $\boldsymbol{c}_{1,k} := (-0.8,-0.3,0.6,4)^\top$ and $\boldsymbol{c}_{2,k} := (-1.4,-0.7,-0.2,0.7,3)^\top$.

B.2.4 Data Analysis: Further Details

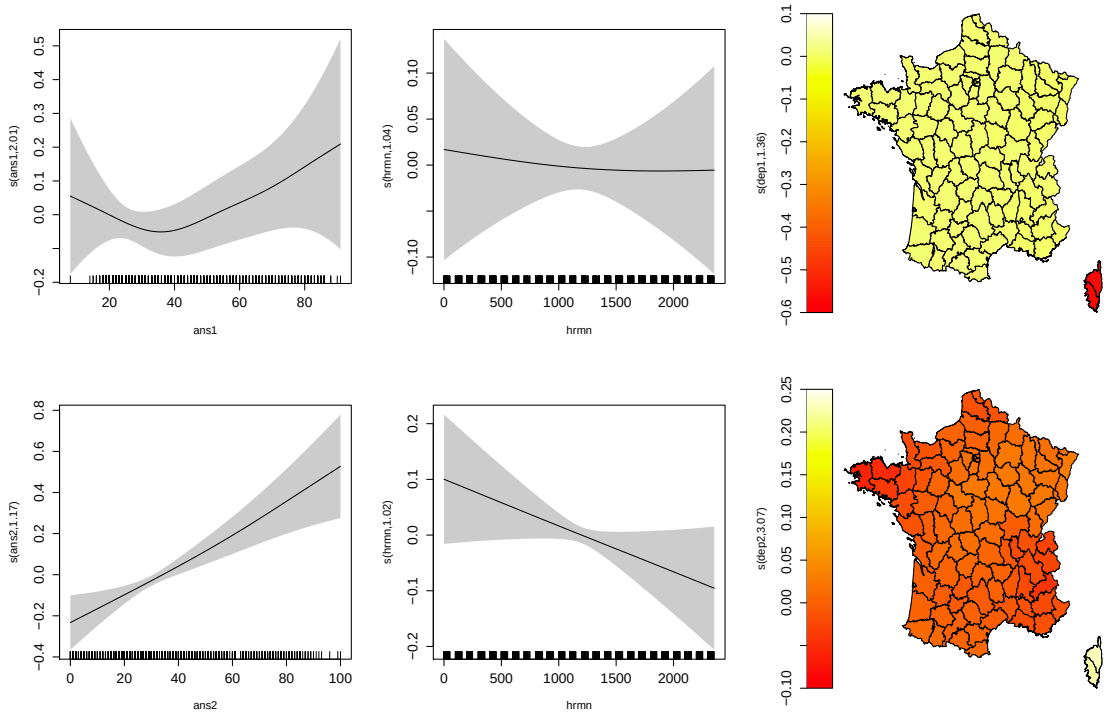

Figure B.5: Smooth functions estimates and associated 95% point-wise confidence intervals corresponding to the two equations (first and second row) of the bivariate model applied to the BAAC 2014 data under Scenario I, using the Joe_0 error dependence. The maps depict graphically the strength of the estimates obtained for the regional variable in every French Departments. We refer to the caption of Figure 3.4 for further details.

Variables	Scenario I: Estimates							
	Joe_0 model				Independent model			
	Driver		Other occupant		Driver		Other occupant	
	estimates	(se)	estimates	(se)	estimates	(se)	estimates	(se)
Occupant Characteristics								
Gender (male)								
female	0.1697	(0.0694)	0.0625	(0.0630)	0.1293	(0.0714)	0.0556	(0.0645)
Seat (other/missing)								
front, passenger	–	–	0.2048	(0.1664)	–	–	0.0665	(0.1687)
rear, driver's side	–	–	0.0364	(0.2344)	–	–	−0.0802	(0.2402)
rear, opposite driver	–	–	−0.0036	(0.2171)	–	–	0.0187	(0.2214)
Motorway Characteristics								
Intersection (off intersection)								
X	0.0395	(0.2358)	−0.2162	(0.2428)	−0.0219	(0.2435)	−0.2347	(0.2467)
T	−0.0924	(0.2535)	−0.3344	(0.2590)	−0.1526	(0.2636)	−0.3319	(0.2611)
Y	−0.2814	(0.3907)	−0.3344	(0.3682)	−0.3370	(0.4052)	−0.3527	(0.3764)
>4 branches	−0.8819	(0.4609)	−0.4818	(0.4177)	−0.9907	(0.4903)	−0.4882	(0.4221)
roundabout	0.3647	(0.2109)	0.3174	(0.2188)	0.3367	(0.2166)	0.3223	(0.2196)
circus/square	0.1333	(1.0256)	−0.4455	(1.0725)	0.1483	(1.0507)	−0.4455	(1.1038)
other	0.3505	(0.2982)	0.2448	(0.3088)	0.3213	(0.3094)	0.2516	(0.3082)
Type (motorway)								
Route Nationale	−0.1418	(0.1333)	−0.0205	(0.1336)	−0.1639	(0.1363)	−0.0415	(0.1359)
Route Départementale	0.0070	(0.0981)	0.1478	(0.0993)	0.0066	(0.1001)	0.1435	(0.1009)
Voie Communale	−0.3588	(0.1056)	−0.0741	(0.1051)	−0.3595	(0.1076)	−0.0712	(0.1073)
other	−0.4516	(0.3301)	−0.3340	(0.3357)	−0.4898	(0.3437)	−0.3417	(0.3387)
Circulation regime (missing)								
one-way	0.0751	(0.1530)	−0.3744	(0.1574)	0.0943	(0.1570)	−0.4131	(0.1584)
two-way	0.3640	(0.1291)	0.1020	(0.1338)	0.3908	(0.1323)	0.0813	(0.1344)
presence of median	0.1197	(0.1430)	−0.1610	(0.1485)	0.1503	(0.1469)	−0.1744	(0.1494)
other	1.0587	(0.5738)	0.1987	(0.5722)	1.0814	(0.5777)	0.0835	(0.5878)
Horizontal alignment (straight)								
left curve	−0.1137	(0.0800)	0.2048	(0.0812)	−0.1311	(0.0816)	0.2102	(0.0820)
right curve	0.0004	(0.0892)	0.0364	(0.0890)	−0.0163	(0.0905)	0.0190	(0.0907)
S	−0.1379	(0.1665)	−0.0036	(0.1657)	−0.0956	(0.1664)	0.0154	(0.1681)
Location (other/missing)								
roadway	−0.4146	(0.1080)	−0.4878	(0.1103)	−0.4539	(0.1095)	−0.5009	(0.1119)
emergency lane	−0.3953	(0.2359)	0.0222	(0.2395)	0.4926	(0.2420)	0.0128	(0.2407)
shoulder	−0.0446	(0.1203)	−0.0932	(0.1221)	−0.0833	(0.1215)	−0.1110	(0.1239)
sidewalk	−0.2552	(0.1864)	−0.6954	(0.1892)	−0.3016	(0.1895)	−0.7241	(0.1924)
Obstacle (other/missing)								
fixed object	0.1938	(0.0887)	0.0984	(0.0904)	0.2044	(0.0907)	0.1057	(0.0915)
pedestrian	−7.2444	$(5.9\,10^5)$	−0.9181	(1.1237)	−7.8729	(3.8175)	−0.8593	(1.1256)
vehicle	−0.1082	(0.1832)	−0.0178	(0.1818)	−0.1191	(0.1883)	−0.0177	(0.1847)
animal	0.1777	(0.3076)	0.0467	(0.3031)	0.2189	(0.3092)	0.0365	(0.3114)
Accident Characteristics								
Lighting (daylight)								
sunrise/sunset	0.0164	(0.1188)	0.1235	(0.1210)	−0.0280	(0.1219)	0.1255	(0.1218)
night without street lights	0.1814	(0.0801)	−0.0430	(0.0788)	0.1591	(0.0824)	−0.0575	(0.0801)
night, street lights in force	−0.0173	(0.0910)	0.1334	(0.0905)	−0.0444	(0.0941)	0.1340	(0.0920)
Atmospheric conditions (normal)								
light rain	−0.2105	(0.0889)	−0.1271	(0.0880)	−0.2097	(0.0904)	−0.1453	(0.0897)
heavy rain	0.4468	(0.1658)	−0.1548	(0.1678)	0.4517	(0.1681)	−0.1743	(0.1714)
snow	0.6575	(0.4330)	−0.6317	(0.4231)	0.8153	(0.4157)	−0.5592	(0.4271)
fog	−0.2960	(0.2864)	0.1544	(0.2890)	−0.2323	(0.2874)	0.1490	(0.2935)
heavy wind/storm	0.0588	(0.4496)	1.1834	(0.4899)	0.0722	(0.4512)	1.1485	(0.4979)
clear	1.0461	(0.4596)	0.6979	(0.4539)	1.0211	(0.4686)	0.6613	(0.4646)
clouds	0.1347	(0.1432)	−0.1661	(0.1458)	0.1352	(0.1461)	−0.1528	(0.1468)
Manner of collision (missing/other)								
heads-on	−0.1864	(0.0761)	0.0176	(0.0773)	−0.1728	(0.0773)	0.0388	(0.0781)
rear-end	−0.3630	(0.1981)	−0.1348	(0.1971)	−0.3834	(0.2054)	−0.1048	(0.1997)
sideswipe, right	−0.1723	(0.1739)	0.7225	(0.1774)	−0.1673	(0.1758)	0.7382	(0.1796)
sideswipe, left	0.6712	(0.1677)	−0.0826	(0.1648)	0.6937	(0.1684)	−0.0902	(0.1682)
Security device (not put on)								
put on	−0.4585	(0.0785)	−0.2646	(0.0771)	−0.4282	(0.0805)	−0.2304	(0.0789)
$c_{j,1}$	−0.9966	(0.2097)	−1.6341	(0.2814)	−0.9943	(0.2142)	−1.6197	(0.2852)
$c_{j,2}$	−0.0786	(0.0196)	−0.2296	(0.0198)	−0.0788	(0.0198)	−0.2111	(0.0199)
$c_{j,3}$	1.3514	(0.0271)	1.3742	(0.0243)	1.3500	(0.0281)	1.3936	(0.0248)
No. observations	1,232		1,232		1,232		1,232	

Table B.5: Estimates and associated standard errors obtained for the parametric model components by applying `CopulaCLM` to the BAAC 2014 data in Scenario I when the Joe_0 copula is used. The last columns report the results corresponding to the independent model. The reference categories are given in round brackets next to the variable names to which they refer.

Variables	Scenario I: Estimates					
	Driver		Other occupant		Independent model	
	estimates	(se)	estimates	(se)	estimates	(se)
Occupant Characteristics						
Gender (male)						
female	0.1697	(0.0694)	0.0625	(0.0630)	0.1094	(0.0466)
Seat (driver)						
other/missing	–	–	ref.	–	0.2970	(0.2347)
front, passenger	–	–	0.2048	(0.1664)	0.3754	(0.0453)
rear, driver's side	–	–	0.0364	(0.2344)	0.1932	(0.1565)
rear, opposite driver	–	–	−0.0036	(0.2171)	0.3175	(0.1333)
Motorway Characteristics						
Intersection (off intersection)						
X	0.0395	(0.2358)	−0.2162	(0.2428)	−0.0927	(0.1714)
T	−0.0924	(0.2535)	−0.3344	(0.2590)	−0.2378	(0.1832)
Y	−0.2814	(0.3907)	−0.3344	(0.3682)	−0.3054	(0.2694)
> 4 branches	−0.8819	(0.4609)	−0.4818	(0.4177)	−0.7115	(0.3076)
roundabout	0.3647	(0.2109)	0.3174	(0.2188)	0.3176	(0.1532)
circus/square	0.1333	(1.0256)	−0.4455	(1.0725)	−0.1332	(0.7553)
other	0.3505	(0.2982)	0.2448	(0.3088)	0.2856	(0.2173)
Type (motorway)						
Route Nationale	−0.1418	(0.1333)	−0.0205	(0.1336)	−0.1005	(0.0952)
Route Départementale	0.0070	(0.0981)	0.1478	(0.0993)	0.0718	(0.0703)
Voie Communale	−0.3588	(0.1056)	−0.0741	(0.1051)	−0.2200	(0.0751)
other	−0.4516	(0.3301)	−0.3340	(0.3357)	−0.4088	(0.2383)
Circulation regime (missing)						
one-way	0.0751	(0.1530)	−0.3744	(0.1574)	−0.1300	(0.1102)
two-way	0.3640	(0.1291)	0.1020	(0.1338)	0.2490	(0.0931)
presence of median	0.1197	(0.1430)	−0.1610	(0.1485)	0.0113	(0.1033)
other	1.0587	(0.5738)	0.1987	(0.5722)	0.6379	(0.4112)
Horizontal alignment (straight)						
left curve	−0.1137	(0.0800)	0.2048	(0.0812)	0.0246	(0.0572)
right curve	0.0004	(0.0892)	0.0364	(0.0890)	−0.0027	(0.0636)
S	−0.1379	(0.1665)	−0.0036	(0.1657)	−0.0397	(0.1175)
Location (other/missing)						
roadway	−0.4146	(0.1080)	−0.4878	(0.1103)	−0.4533	(0.0775)
emergency lane	−0.3953	(0.2359)	0.0222	(0.2395)	−0.2426	(0.1688)
shoulder	−0.0446	(0.1203)	−0.0932	(0.1221)	−0.0861	(0.0860)
sidewalk	−0.2552	(0.1864)	−0.6954	(0.1892)	−0.4705	(0.1336)
Obstacle (other/missing)						
fixed object	0.1938	(0.0887)	0.0984	(0.0904)	0.1594	(0.0636)
pedestrian	−7.2444	$(5.9\ 10^5)$	−0.9181	(1.1237)	−1.0730	(0.8588)
vehicle	−0.1082	(0.1832)	−0.0178	(0.1818)	−0.0639	(0.1301)
animal	0.1777	(0.3076)	0.0467	(0.3031)	0.1418	(0.2180)
Accident Characteristics						
Lighting (daylight)						
sunrise/sunset	0.0164	(0.1188)	0.1235	(0.1210)	0.0424	(0.0853)
night without street lights	0.1814	(0.0801)	−0.0430	(0.0788)	0.0709	(0.0567)
night, street lights in force	−0.0173	(0.0910)	0.1334	(0.0905)	0.0420	(0.0648)
Atmospheric conditions (normal)						
light rain	−0.2105	(0.0889)	−0.1271	(0.0880)	−0.1761	(0.0630)
heavy rain	0.4468	(0.1658)	−0.1548	(0.1678)	0.1702	(0.1189)
snow	0.6575	(0.4330)	−0.6317	(0.4231)	0.1938	(0.2940)
fog	−0.2960	(0.2864)	0.1544	(0.2890)	−0.0662	(0.2038)
heavy wind/storm	0.0588	(0.4496)	1.1834	(0.4899)	0.4839	(0.3227)
clear	1.0461	(0.4596)	0.6979	(0.4539)	0.8283	(0.3260)
clouds	0.1347	(0.1432)	−0.1661	(0.1458)	−0.0031	(0.1027)
Manner of collision (missing/other)						
head-on	−0.1864	(0.0761)	0.0176	(0.0773)	−0.0831	(0.0544)
rear-end	−0.3630	(0.1981)	−0.1348	(0.1971)	−0.2355	(0.1411)
sideswipe, right	−0.1723	(0.1739)	0.7225	(0.1774)	0.2286	(0.1237)
sideswipe, left	0.6712	(0.1677)	−0.0826	(0.1648)	0.3176	(0.1178)
Security device (not put on)						
put on	−0.4585	(0.0785)	−0.2646	(0.0771)	−0.3452	(0.0554)
$c_{j,1}$	−0.9966	(0.2097)	−1.6341	(0.2814)	−1.0693	(0.1515)
$c_{j,2}$	−0.0786	(0.0196)	−0.2296	(0.0198)	0.0378	(0.0138)
$c_{j,3}$	1.3514	(0.0271)	1.3742	(0.0243)	1.5235	(0.0180)
No. observations	1,232		1,232		2,464	

Table B.6: Estimates for Scenario I: the independent model is obtained under a univariate model where all the observations are pooled together.

Variables	Scenario I: Pseudo-elasticities					
	Joe_0: Semi-parametric		Independent		Joe_0: Parametric	
	Driver	Other occupant	Driver	Other occupant	Driver	Other occupant
Occupant Characteristics						
Gender (male)						
female	15.1771	4.4118	11.4417	3.9410	20.0326	6.3791
Seat (other/missing)						
front, passenger	–	3.3171	–	5.0900	–	11.2427
rear, driver's side	–	−0.6495	–	−5.2680	–	−2.2810
rear, opposite driver	–	2.5478	–	1.3643	–	10.9453
Motorway Characteristics						
Intersection (off intersection)						
X	3.5046	−15.5834	−1.9336	−16.9575	20.8335	−8.8219
T	−8.1106	−24.0573	−13.2887	−23.9279	0.2473	−24.1377
Y	−24.0709	−24.0592	−28.5440	−25.4054	−6.8963	−11.5996
$>$ 4 branches	−64.7161	−34.3071	−70.0681	−34.7874	−56.1683	−38.8870
roundabout	32.9388	21.1949	29.4565	21.5968	40.6659	11.2428
circus/square	11.9005	−31.8270	13.1213	−31.8771	21.0943	16.2180
other	31.6143	16.6715	28.16066	17.1848	34.1223	23.2436
Type (motorway)						
Route Nationale	−12.3814	−1.4598	−14.2561	−2.9740	−7.4069	−0.0587
Route Départementale	0.6239	10.2816	0.5852	10.0325	3.5596	12.7703
Voie Communale	−30.2670	−5.3074	−30.3170	−5.1137	−20.9363	−2.6323
other	−37.3652	−24.0312	−40.1773	−24.6226	−31.9717	−30.2226
Circulation regime (missing)						
one-way	6.6856	−26.8814	8.3478	−29.6338	15.9702	−25.6134
two-way	32.8730	7.1588	33.9659	5.7432	36.7234	11.4425
presence of median	10.6766	−11.5934	13.2971	−12.5896	20.2130	−10.2389
other	285.5734	13.6792	73.6875	5.8991	69.6891	13.6003
Horizontal alignment (straight)						
left curve	−8.9555	18.0840	−10.1281	18.7363	−7.1916	15.3640
right curve	0.0402	2.5769	−1.4410	1.3507	−3.3128	−0.5520
S	−10.5973	−0.2560	−7.6513	1.1204	−9.4537	6.7268
Location (other/missing)						
roadway	−34.5838	−34.7099	−37.5420	−35.6480	−33.3844	−35.0592
emergency lane	−33.1021	1.5788	−40.3814	0.9101	−38.7264	−1.8851
shoulder	−3.9319	−6.6893	−7.3046	−7.9924	−9.8526	−9.4138
sidewalk	−21.9244	−48.0965	−25.7039	−49.8746	−23.9135	−51.1506
Obstacle (other/missing)						
fixed object	21.1914	7.7640	22.3652	8.4292	22.3471	8.8375
pedestrian	–	−21.5275	–	−22.3293	–	−22.2966
vehicle	−8.5727	−1.2459	−9.3135	−1.2383	−7.4586	9.0969
animal	19.0619	3.4888	24.3422	2.7093	6.4797	10.0389
Accident Characteristics						
Lighting (daylight)						
sunrise/sunset	1.4600	8.6284	−2.4667	8.7987	−0.4411	6.5159
night without street lights	16.2326	−3.0728	14.0798	−4.1294	20.2537	−5.9967
night, street lights in force	−1.5261	9.3080	3.9102	9.3804	−2.7762	10.9983
Atmospheric conditions (normal)						
light rain	−18.2084	−9.1396	−18.1288	−10.4779	−15.9728	−5.1552
heavy rain	40.7718	−11.1393	38.8988	−12.5850	55.3536	−12.7853
snow	65.8266	−44.1370	63.4047	−39.5186	58.4414	−59.4227
fog	−19.4357	12.9200	−16.3039	12.4447	−17.2725	10.5198
heavy wind/storm	5.2257	45.3357	6.38444	46.4522	−6.1463	53.2024
clear	267.3856	39.6371	72.0389	38.6520	121.7924	52.1828
clouds	12.0293	−11.9607	11.9639	−11.0216	20.5940	−11.6020
Manner of collision (missing/other)						
heads-on	−16.1804	1.2487	−15.0135	2.7577	−22.0029	2.4676
rear-end	−22.3114	−8.3346	−23.1520	−6.7087	−26.7919	−9.9022
sideswipe, right	−12.7931	116.2993	−12.4661	121.8290	−20.3684	129.3245
sideswipe, left	68.0130	−5.9216	56.3359	−6.4898	60.6336	−9.6748
Security device (not put on)						
put on	−25.6391	−14.2845	−24.7310	−12.9336	−27.2415	−14.1851
No. observations	1,232		1,232		1,232	

Table B.7: Pseudo-elasticities of the parametric model components of Scenario I obtained by applying the preferred Joe_0 copula, independent and the purely parametric models. The reported quantities are computed with respect to the hospitalised injuries.

References

Abay, K., Paleti, R., and Bhat, C. (2013). The joint analysis of injury severity of drivers in two-vehicle crashes accommodating seat belt use endogeneity. *Transportation Research Part B*, 50:74–89. (Cited on page 55.)

Aitchison, J. and Silvey, S. (1957). The generalization of probit analysis to the case of multiple responses. *Biometrika*, 44(1/2):131–140. (Cited on page 19.)

Anderson, J. and Philips, P. (1981). Regression, discrimination and measurement models for ordered categorical variables. *Journal of the Royal Statistical Society, Series C*, 1(1):22–31. (Cited on page 20.)

Angris, J., Imbens, G., and Krueger, A. (1999). Jackknife Instrumental Variables estimation. *Journal of Applied Econometrics*, 14(1):57–67. (Cited on page 16.)

Angris, J. and Krueger, A. (2001). Instrumental variables and the search for identification: From supply and demand to natural experiments. *Journal of Economic Perspectives*, 15(4):69–85. (Cited on page 11.)

Arpino, B., De Cao, E., and Peracchi, F. (2014). Using panel data for partial identification of Human Immunodeficiency Virus prevalence when infection status is missing not at random. *Journal of the Royal Statistical Society, Series A*, 177(3):587–606. (Cited on pages 90 and 93.)

Barndorff-Nielsen, O. and Cox, D. (1994). *Inference and Asymptotics*. Chapman & Hall, London, UK. (Cited on page 97.)

Bärnighausen, T., Bor, J., Wandira-Kazibwe, S., and Canning, D. (2011). Correcting HIV prevalence estimates for survey nonparticipation using Heckman-type selection models. *Epidemiology*, 22(1):27–35. (Cited on page 91.)

Becher, H. (1992). The concept of residual confounding in regression models and some applications. *Statistics in Medicine*, 11(13):1747–1758. (Cited on page 77.)

Bhat, C. and Eluru, N. (2009). A copula-based approach to accommodate residential self-selection effects in travel behavior modeling. *Transportation Research Part B*, 43(7):749–765. (Cited on page 49.)

Blundell, R., Dearden, L., Goodman, D., and Reed, H. (2000). The returns to Higher Education in Britain: Evidence from a british cohort. *The Economic Journal*, 110(461):F82–F99. (Cited on page 45.)

Boerma, J., Ghys, P., and Walker, N. (2003). Estimates of HIV-1 prevalence from national population-based surveys as a new gold standard. *The Lancet*, 362(9399):1929–1931. (Cited on page 90.)

Bound, J., Jaeger, D., and Baker, R. (1995). Problems with instrumental variables estimation when the correlation between the instruments and the endogeneous explanatory variable is weak. *Journal of the American Statistical Association*, 90(430):443–450. (Cited on page 15.)

Bratti, M. and Miranda, A. (2009). Selection-endogenous ordered probit and dynamic ordered probit models. *Proceedings of the United Kingdom Stata Users' Group Meetings 2009*. (Cited on pages 40 and 41.)

Bratti, M. and Miranda, A. (2010). Non-pecuniary returns to Higher Education: The effect on smoking intensity in the UK. *Health Economics*, 19(8):906–920. (Cited on pages 39 and 41.)

Brechmann, E. and Schepsmeier, U. (2013). Modeling dependence with C- and D-vine copulas: The R package CDVine. *Journal of Statistical Software*, 52(3). (Cited on pages x, 52 and 54.)

Brunello, G., Michaud, P., and Sanz-de Galdeano, A. (2008). The rise in obesity across the Atlantic: An economic perspective. *IZA Discussion Paper No. 3529*. (Cited on page 38.)

Buscha, F. and Conte, A. (2014). The impact of truancy on educational attainment during compulsory schooling: A bivariate ordered probit estimator with mixed effects. *The Manchester School*, 82(1):103–127. (Cited on pages 21, 26 and 45.)

Caldwell, T., Rodgers, B., Clark, C., Jefferis, B., Stansfeld, S., and Power, C. (2008). Lifecourse socioeconomic predictors of midlife drinking patterns, problems and abstention: Findings from the 1958 British Birth Cohort Study. *Drug and Alcohol Dependence*, 95(3):269–278. (Cited on page 40.)

Chib, S. and Greenberg, E. (2007). Semiparametric modeling and estimation of instrumental

variable models. *Journal of Computational and Graphical Statistics*, 16(1):86–114. (Cited on page 110.)

Chiou, Y.-C., Hwang, C.-C., Chang, C.-C., and Fu, C. (2013). Modeling two-vehicle crash severity by a bivariate generalized ordered probit approach. *Accident Analysis and Prevention*, 51:175–184. (Cited on page 64.)

Cox, D. and Wermuth, N. (1993). Linear dependencies represented by chain graphs. *Statistical Science*, 8(3):204–218. (Cited on page 13.)

Cox, D. and Wermuth, N. (2003). A general condition for avoiding effect reversal after marginalization. *Journal of the Royal Statistical Society, Series B*, 65(4):937–941. (Cited on page 12.)

Cox, D. and Wermuth, N. (2004). Causality: A statistical view. *International Statistical Review*, 72(3):285–305. (Cited on pages 14, 15 and 26.)

Craven, P. and Wahba, G. (1979). Smoothing noisy data with spline functions. Estimating the correct degree of smoothing by the method of Generalized Cross-Validation. *Numerische Mathematik*, 31(4):377–403. (Cited on page 32.)

Dale, J. (1986). Global cross-ratio models for bivariate, discrete, ordered responses. *Biometrics*, 42(4):909–917. (Cited on page 24.)

Davey, B. and Priestley, H. (2002). *Introduction to Lattices and Order*. Cambridge University Press, New York, NY. (Cited on page 100.)

de Boor, C. (1978). *A Practical Guide to Splines*. Springer, New York, NY. (Cited on page 7.)

de Lapparent, M. (2008). Willingness to use safety belts and levels of injury in car accidents. *Accident Analysis and Prevention*, 40(3):1023–1032. (Cited on page 75.)

Delaney, L., Harmon, C., and Wall, P. (2008). Behavioral economics and drinking behavior: Preliminary results from an Irish college study. *Economic Inquiry*, 46(1):269–272. (Cited on page 39.)

Didelez, V., Meng, S., and Sheehan, N. (2010). Assumptions of IV methods for observational epidemiology. *Statistical Science*, 25(1):22–40. (Cited on page 14.)

Didelez, V. and Sheehan, N. (2007). Mendelian randomization as an instrumental variable approach to causal inference. *Statistical Methods in Medical Research*, 16(4):309–330. (Cited on page 14.)

Droomers, M., Schrijvers, C., Casswell, S., and Mackenbach, J. (2003). Occupational level of the father and alcohol consumption during adolescence; patterns and predictors. *Journal of Epidemiology and Community Health*, 57(9):704–710. (Cited on page 41.)

Duchon, J. (1977). *Construction Theory of Functions of Several Variables*, chapter Splines Minimizing Rotation-invariant Semi-norms in Solobev Spaces, pages 85–100. Springer, Berlin. (Cited on page 8.)

Durante, F. (2009). Construction of non-exchangeable bivariate distribution functions. *Statistical Papers*, 50(2):383–391. (Cited on page 53.)

Eilers, P. and Marx, B. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, 11(2):89–121. (Cited on pages 7, 25, 55 and 79.)

Eluru, N., Paleti, R., Pendyala, R., and Bhat, C. (2010). Modeling injury severity of multiple occupants of vehicles: Copula-based multivariate approach. *Transportation Research Record: Journal of the Transportation Research Board*, 2165:1–11. (Cited on pages 49, 64, 67 and 72.)

Fahrmeir, L. and Tutz, G. (2001). *Multivariate Statistical Modelling Based on Generalized Linear Models*. Springer, New York, NY. (Cited on page 20.)

Fehr, E. (2002). The economics of impatience. *Nature*, 415(6869):269–272. (Cited on page 39.)

Frees, E. and Valdez, E. (1998). Understanding relationships using copulas. *North American Actuarial Journal*, 2(1):1–25. (Cited on page 53.)

Friedman, J. and Silverman, B. (1989). Flexible parsimonious smoothing and additive modeling. *Technometrics*, 31(1):3–21. (Cited on page 6.)

Frosini, B. (2006). Causality and causal models: A conceptual perspective. *International Statistical Review*, 74(3):305–334. (Cited on pages 11 and 26.)

Fuchs, V. (1982). *Economic Aspects of Health*, chapter Time Preference and Health: An Exploratory Study. University of Chicago Press, Chicago, IL. (Cited on page 39.)

Fuller, W. (1977). Some properties of a modification of the limited information estimator. *Econometrica*, 45(4):939–953. (Cited on page 16.)

Gersovitz, M. (2011). HIV testing: Principles and practice. *The World Bank Research Observer*, 26(1):1–41. (Cited on page 90.)

Gertheiss, J. and Tutz, G. (2009). Penalized regression with ordinal predictors. *International Statistical Review*, 77(3):345–365. (Cited on page 46.)

Geyer, C. (2013). Trust regions. `http://cran.stat.ucla.edu/web/packages/trust/vignettes/trust.pdf`. (Cited on pages 32 and 59.)

Goldman, D. and Smith, J. (2005). Socioeconomic differences in the adoption of new medical technologies. *American Economic Review*, 95(2):234–237. (Cited on page 38.)

Green, P. (1984). Iteratively Reweighted Least Squares for Maximum Likelihood estimation, and some robust and resistant alternatives (with discussion). *Journal of the Royal Statistical Society, Series B*, 46(2):149–192. (Cited on pages 33 and 88.)

Green, P. and Silverman, B. (1994). *Nonparametric Regression and Generalized Linear Models. A Roughness Penalty Approach.* Chapman & Hall, London, UK. (Cited on pages 30 and 57.)

Greene, W. and Hensher, D. (2010). *Modeling Ordered Choices. A Primer.* Cambridge University Press, Cambridge, UK. (Cited on pages 20, 26 and 44.)

Haberman, S. (1980). Discussion of McCullagh (1980). *Journal of the Royal Statistical Society, Series B*, 42(2):136–137. (Cited on page 24.)

Hastie, T. and Tibshirani, R. (1986). Generalized Additive Models (with discussion). *Statistical Science*, 1(3):297–318. (Cited on pages 1, 49 and 77.)

Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models.* Chapman & Hall, London, UK. (Cited on pages 1, 49 and 77.)

Heckman, J. (1978). Dummy endogenous variables in a simultaneous equation system. *Econometrica*, 46(4):931–959. (Cited on pages 26 and 86.)

Heckman, J. (1979). Sample selection bias as a specification error. *Econometrica*, 47(1):153–161. (Cited on pages 86 and 87.)

Hemmingsson, T., Lundberg, I., and Diderichsen, F. (1999). The roles of social class of origin, achieved social class and intergenerational social mobility in explaining social class inequalities in alcoholism among young men. *Social Science & Medicine*, 49(8):1051–1059. (Cited on page 41.)

Hillmann, J., Kneib, T., Koepcke, L., Paz, L., and Kretzberg, J. (2014). Bivariate cumulative probit model for the comparison of neuronal encoding hypotheses. *Biometrical Journal*, 56(1):23–43. (Cited on pages 21, 49 and 72.)

Hogan, D., Salomon, J., Canning, D., Hammitt, J., Zaslavsky, A., and Bärnighausen, T. (2012). National HIV prevalence estimates for Sub-Saharan Africa: Controlling selection bias with Heckman-type selection models. *Sexually Transmitted Infections*, 88:i17–i23. (Cited on pages 90 and 91.)

Horowitz, J. and Manski, C. (2000). Nonparametric analysis of randomized experiments with missing covariate and outcome data. *Journal of the American Statistical Association*, 95(449):77–84. (Cited on page 93.)

Huerta, M. and Borgonovi, F. (2010). Education, alcohol use and abuse among young adults in Britain. *Social Science & Medicine*, 71(1):143–151. (Cited on pages 38 and 40.)

Imbens, G. (2014). Instrumental variables: An econometrician's perspective. *Statistical Science*, 29(3):323–358. (Cited on page 86.)

Janssens, W., van der Gaag, J., de Wit, T., and Tanović, Z. (2014). Refusal bias in the estimation of HIV prevalence. *Demography*, 51(3):1131–1157. (Cited on page 91.)

Kauermann, G. (2005). Penalized spline smoothing in multivariable survival models with varying coefficients. *Computational Statistics & Data Analysis*, 49(1):169–186. (Cited on pages 36, 37 and 97.)

Kauermann, G., Krivobokova, T., and Fahrmeir, L. (2009). Some asymptotic results on generalized penalized spline smoothing. *Journal of the Royal Statistical Society, Series B*, 71(2):487–503. (Cited on page 98.)

Kawakatsu, H. and Largey, A. (2009). EM algorithms for ordered probit models with endogenous regressors. *The Econometrics Journal*, 12(1):164–186. (Cited on page 56.)

Keane, M. (1992). A note on identification in the multinomial probit model. *Journal of Business & Economic Statistics*, 10(2):193–200. (Cited on page 31.)

Kenkel, D. (1991). Health behavior, health knowledge, and schooling. *Journal of Political Economy*, 99(2):287–305. (Cited on page 38.)

Kim, K. (1995). A bivariate cumulative probit regression model for ordered categorical data. *Statistics in Medicine*, 14(12):337–356. (Cited on page 28.)

Kim, Y. and Gu, C. (2004). Smoothing spline Gaussian regression: More scalable computation via efficient approximation. *Journal of the Royal Statistical Society, Series B*, 66(2):337–356. (Cited on page 35.)

Klein, N. and Kneib, T. (2015). Simultaneous inference in structured additive conditional copula regression models: A unifying Bayesian approach. *Statistics and Computing (in press)*. (Cited on pages 22 and 79.)

Klein, N., Kneib, T., Klasen, S., and Lang, L. (2015). Bayesian structured additive distributional regression for multivariate responses. *Journal of the Royal Statistical Society, Series C*, 64(4):569–591. (Cited on pages xvi, 22, 77 and 81.)

Kneib, T. (2005). *Mixed Model Based Inference in Structured Additive Regression*. PhD Thesis, Ludwig-Maximilians-Universität München, München, Germany. (Cited on page 58.)

Kneib, T. (2013). Beyond mean regression. *Statistical Modelling*, 13(4):275–303. (Cited on page 77.)

Kosmidis, I. (2014). Improved estimation in Cumulative Link Models. *Journal of the Royal Statistical Society, Series B*, 76(1):169–196. (Cited on page 24.)

Li, X., Lord, D., and Zhang, Y. (2011). Development of accident modification factors for rural frontage road segments in texas using generalized additive models. *Journal of Transportation Engineering*, 137(1):74–83. (Cited on page 49.)

Mannering, F. (2009). An empirical analysis of driver perceptions of the relationship between speed limits and safety. *Transportation Research Part F*, 12(2):99–106. (Cited on page 69.)

Mannering, F. and Bhat, C. (2014). Analytic methods in accident research: Methodological frontier and future directions. *Analytic Methods in Accident Research*, 1:1–22. (Cited on pages 49 and 55.)

Manski, C. F. (1995). *Identification Problems in the Social Sciences*. Harvard University Press, Cambridge, MA. (Cited on page 93.)

Manski, C. F. (2003). *Partial Identification of Probability Distributions.* Springer-Verlag, New York, NY. (Cited on page 93.)

Marra, G. and Radice, R. (2011). Estimation of a semiparametric recursive bivariate probit model in the presence of endogeneity. *The Canadian Journal of Statistics*, 39(2):259–279. (Cited on pages 75, 87 and 110.)

Marra, G. and Radice, R. (2013). A Penalized Likelihood estimation approach to semiparametric sample selection binary response modeling. *The Electronic Journal of Statistics*, 7:1432–1455. (Cited on pages 31 and 87.)

Marra, G. and Radice, R. (2015). *SemiParBIVProbit: Semiparametric Bivariate Probit Modelling.* R package version 3.3. (Cited on page 89.)

Marra, G., Radice, R., Bärnighausen, T., Wood, S., and McGovern, M. (2015). A unified modeling approach to estimating HIV prevalence in Sub-Saharan African countries. *Research Report No. 324, Department of Statistical Science, University College London.* (Cited on pages 17, 60, 66 and 89.)

Marra, G. and Wood, S. (2011). Practical variable selection for Generalized Additive Models. *Computational Statistics & Data Analysis*, 55(7):2372–2387. (Cited on page 43.)

Marra, G. and Wood, S. (2012). Coverage properties of confidence intervals for Generalized Additive Model components. *Scandinavian Journal of Statistics*, 39(1):53–74. (Cited on pages x, 35, 63 and 70.)

McCullagh, P. (1980). Regression models for ordinal data (with discussion). *Journal of the Royal Statistical Society, Series B*, 42(2):109–142. (Cited on pages 4, 19, 48, 53 and 83.)

McGovern, M., Bärnighausen, T., Marra, G., and Radice, R. (2015). On the assumption of bivariate normality in selection models. A copula approach applied to estimating HIV prevalence. *Epidemiology*, 26(2):229–237. (Cited on pages 89, 90, 91 and 92.)

McKelvey, R. and Zavoina, W. (1975). A statistical model for the analysis of ordinal level dependent variables. *The Journal of Mathematical Sociology*, 4(1):109–142. (Cited on pages 20, 27, 53 and 84.)

McNeil, A. and Nešlehová, J. (2009). Multivariate Archimedean copulas, d-monotone functions and ℓ_1-norm symmetric distributions. *The Annals of Statistics*, 37(5B):3059–3097. (Cited on page 52.)

Miranda, A. and Rabe-Hesketh, S. (2006). Maximum Likelihood Estimation of endogenous switching and sample selection models for binary, ordinal, and count variables. *The Stata Journal*, 6(3):285–308. (Cited on page 87.)

Montana, L., Mishra, V., and Hong, R. (2008). Measuring the HIV/AIDS epidemic: Approaches and challenges. *Sexually Transmitted Infections*, 84:i78–i84. (Cited on page 90.)

Nelder, J. and Wedderburn, R. (1972). Generalized Linear Models. *Journal of the Royal Statistical Society, Series A*, 135(3):370–384. (Cited on pages 2, 19, 49 and 76.)

Nelsen, R. (2006). *An Introduction to Copulas*. Springer, New York, NY. (Cited on page 53.)

Nocedal, J. and Wright, S. (2006). *Numerical Optimization*. Springer, New York, NY. (Cited on pages 32 and 59.)

O'Donoghue, T. and Rabin, M. (2000). The economics of immediate gratification. *Journal of Behavioural Decision Making*, 13(2):233–250. (Cited on page 39.)

OECD (2014). *Health at a Glance: Europe 2014*. OECD Publishing. (Cited on page 38.)

O'Sullivan, F., Yandell, B., and Raynor, W. (1986). Automatic smoothing of regression functions in Generalized Linear Models. *Journal of the American Statistical Association*, 81(393):96–103. (Cited on pages 30, 34 and 59.)

Parker, R. and Rice, J. (1985). Discussion of silverman (1985). *Journal of the Royal Statistical Society, Series B*, 47(1):40–42. (Cited on page 7.)

Peyhardi, J., Trottier, C., and Guédon, Y. (2014). A new specification of Generalized Linear Models for categorical data. *arXiv:1404.7331v2*. (Cited on pages 3, 4, 21, 22, 50, 57, 78 and 93.)

Poulton, R., Caspi, A., Milne, B., Murray Thomson, W., Taylor, A., Sears, M., and Moffitt, T. (2002). Association between children's experience of socioeconomic disadvantage and adult health: A life-course study. *The Lancet*, 360(9346):1640–1645. (Cited on page 41.)

Pratt, J. (1981). Concavity of the log likelihood. *Journal of the American Statistical Association*, 76(373):103–106. (Cited on page 24.)

Public Health England (2014). *Alcohol Treatment in England 2013-2014*. Public Health England, London, UK. (Cited on page 38.)

R Development Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. (Cited on pages 49 and 89.)

Radice, R., Marra, G., and Wojtyś, M. (2015). Copula regression spline models for binary outcomes. *Statistics and Computing (in press)*. (Cited on pages 21, 31, 48, 49, 66, 87 and 92.)

Rana, T., Sikder, S., and Pinjari, A. (2010). Copula-based method for addressing endogeneity in models of severity of traffic crash injuries. *Transportation Research Record: Journal of the Transportation Research Board*, 2147:75–87. (Cited on page 49.)

Rigby, R. and Stasinopoulos, D. (2005). Generalized Additive Models for Location, Scale and Shape. *Journal of the Royal Statistical Society, Series C*, 54(3):507–554. (Cited on page 77.)

Rousseeuw, P. J. (1984). Least median of squares regression. *Journal of the American Statistical Association*, 79(388):871–880. (Cited on page 96.)

Royston, P. and Altman, D. (1994). Regression using fractional polynomials of continuous covariates: Parsimonious parametric modelling (with discussion). *Journal of the Royal Statistical Society, Series C*, 43(3):429–467. (Cited on page 21.)

Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields*. Chapman & Hall/CRC, Boca Raton, FL. (Cited on pages xvi, 58 and 81.)

Ruppert, D., Wand, M., and Carroll, R. (2003). *Semiparametric Regression*. Cambridge University Press, Cambridge, UK. (Cited on pages 25, 55 and 80.)

Russo, B., Savolainen, P., Schneider IV, W., and Anastasopoulos, P. (2014). Comparison of factors affecting injury severity in angle collisions by fault status using a random parameters bivariate ordered probit model. *Analytic Methods in Accident Research*, 2:21–29. (Cited on pages 63 and 64.)

Sajaia, Z. (2008). Maximum Likelihood Estimation of a bivariate ordered probit model: Implementation and monte carlo simulations. *Unpublished manuscript*. (Cited on pages 21, 26, 29 and 45.)

Sander, W. (1995). Schooling and quitting smoking. *The Review of Economics and Statistics*, 77(1):191–199. (Cited on page 39.)

Silverman, B. (1985). Some aspects of the spline smoothing approach to non-parametric regression curve fitting (with discussion). *Journal of the Royal Statistical Society, Series B*, 47(1):1–52. (Cited on page 35.)

Sklar, A. (1959). Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institut de Statistique de l'Université de Paris*, 8:229–231. (Cited on pages 51 and 79.)

Snell, E. (1964). A scaling procedure for ordered categorical data. *Biometrics*, 20(3):592–607. (Cited on page 19.)

StataCorp (2015). *STATA: Data Analysis and Statistical Software: Release 13.* (Cited on page 21.)

Sterck, O. (2013). Why are testing rates so low in Sub-Saharan Africa? Misconceptions and strategic behaviors. *Forum for Health Economics & Policy*, 16(1):219–257. (Cited on page 90.)

Stevens, S. (1946). On the theory of scales of measurement. *Science*, 103(2684):677–680. (Cited on pages 3 and 78.)

Stock, J., Wright, J., and Yogo, M. (2002). A survey of weak instruments and weak identification in generalized method of moments. *Journal of Business & Economic Statistics*, 20(4):518–529. (Cited on page 15.)

Stone, C., Hansen, M., Kooperberg, C., and Truong, Y. (1997). Polynomial splines and their tensor products in extended linear modeling. *The Annals of Statistics*, 25(4):1371–1425. (Cited on page 6.)

Trivedi, P. and Zimmer, D. (2005). Copula modeling: An introduction for practitioners. *Foundations and Trends in Econometrics*, 1(1):1–111. (Cited on pages 53, 62 and 66.)

UCL Institute of Education. Centre for Longitudinal Studies (2007). *Millennium Cohort Study: First Survey, 2001-2003 [computer file].* UK Data Archive [distributor], Colchester, Essex, UK. (Cited on page 22.)

Ulfarsson, G. and Mannering, F. (2013). Differences in male and female injury severities in Sport-Utility Vehicles, minivan. pickup and passenger car accidents. *Accident Analysis and Prevention*, 36(2):135–147. (Cited on page 67.)

UNAIDS-World Health Organization (2007). *Guidelines for Conducting HIV Sentinel Serosurveys among Pregnant Women and Other Groups*. UNAIDS, Geneva, CH. (Cited on page 90.)

van der Pol, M. (2011). Health, education and time preference. *Health Economics*, 20(8):906–920. (Cited on page 39.)

Vossmeyer, A. (2014). Determining the proper specification for endogenous covariates in discrete data settings. *Advances in Econometrics*, 34:223–247. (Cited on pages 86 and 87.)

Wahba, G. (1980). *Approximation Theory III*, chapter Spline Bases, Regularization, and Generalized Cross Validation for Solving Approximation Problems with Large Quantities of Noisy Data. Academic Press, London, UK. (Cited on page 7.)

Wahba, G. (1983). Bayesian "confidence intervals" for the cross-validated smoothing spline. *Journal of the Royal Statistical Society, Series B*, 45(1):133–150. (Cited on page 35.)

Wermuth, N. and Cox, D. (2008). Distortion of effects caused by indirect confounding. *Biometrika*, 98(1):481–493. (Cited on pages 12 and 26.)

Wood, S. (2003). Thin plate regression splines. *Journal of the Royal Statistical Society, Series B*, 65(1):481–493. (Cited on pages xii, 8, 9, 10, 25, 55, 80 and 92.)

Wood, S. (2004). Stable and efficient multiple smoothing parameter estimation for Generalized Additive Models. *Journal of the American Statistical Association*, 99(467):673–686. (Cited on pages 1, 32 and 61.)

Wood, S. (2006). *Generalized Additive Models. An Introduction With R*. Chapman & Hall/CRC, Boca Raton, FL. (Cited on pages xvi, 9, 25, 30, 34, 36, 54, 55 and 81.)

Wooldridge, J. (2002). *Econometric Analysis of Cross Section and Panel Data*. MIT Press, Cambridge, MA. (Cited on pages 11, 13 and 85.)

World Health Organization (2004). *World Report on Road Traffic Injury Prevention*. (Cited on page 48.)

World Health Organization (2007). *A60/14 Add.1, 60th World Health Assembly, Provisional Agenda Item 12.7*. (Cited on page 38.)

World Health Organization (2013). *Global Status Report on Road Safety 2013: Supporting a Decade of Action*. (Cited on page 47.)

Xie, Y. and Zhang, Y. (2008). Crash frequency analysis with generalized additive models. *Transportation Research Record: Journal of the Transportation Research Board*, 2061:39–45. (Cited on page 49.)

Yamamoto, T. and Shankar, V. (2004). Bivariate ordered-response probit model of driver's and passenger's injury severities in collisions with fixed objects. *Accident Analysis and Prevention*, 36(5):869–876. (Cited on pages 28 and 64.)

Yee, T. and Wild, C. (1996). Vector Generalized Additive Models. *Journal of the Royal Statistical Society, Series B*, 58(3):481–493. (Cited on pages 1, 22, 29, 49, 60 and 93.)

Zhang, Q. and Ip, E. (2012). Generalized Linear Model for partially ordered data. *Statistics in Medicine*, 31(1):56–68. (Cited on page 22.)